Textual Analysis in Accounting and Finance: A Survey

Tim Loughran
Mendoza College of Business
University of Notre Dame
Notre Dame, IN 46556-5646
574.631.8432 *voice*
Loughran.9@nd.edu

Bill McDonald
Mendoza College of Business
University of Notre Dame
Notre Dame, IN 46556-5646
574.631.5137 *voice*
mcdonald.1@nd.edu

May 20, 2016

ABSTRACT

Relative to quantitative methods traditionally used in accounting and finance, textual analysis is substantially less precise. Thus, understanding the art is of equal importance to understanding the science. In this survey we describe the nuances of the method and, as users of textual analysis, some of the tripwires in implementation. We also review the contemporary textual analysis literature and highlight areas of future research.

Keywords: Textual analysis; sentiment analysis; bag-of-words; readability; word lists; Zipf's law; cosine similarity; Naïve Bayes

JEL Codes: D82; D83; G14; G18; G30; M40; M41

Accepted by Christian Leuz. We thank Brad Badertscher, Peter Easton, Diego Garcia, two anonymous referees, and seminar participants at Columbia Business School's News and Finance Conference for helpful comments.

1. Introduction

Textual analysis, in some form, resides across many disciplines under various aliases including computational linguistics, natural (or statistical) language processing, information retrieval, content analysis, or stylometrics. The notion of parsing text for patterns has a long history. In the 1300's, friars of the Dominican order produced concordances of the Latin Vulgate (Biblical translations) to provide indexes of common phrases (Catholic Encyclopedia, Vol. 4, 1908). In 1901, T.C. Mendenhall used textual analysis to examine whether some works attributed to Shakespeare might have been written by Bacon (see Williams [1975]). During the world wars, the method was increasingly adapted to political speech, where carefully scripted rhetorical choices were interpreted as signals of diplomatic trends (e.g., Burke [1939]). In the sixties, the systematic analysis of text increased in popularity with Mosteller and Wallace's [1964] purported resolution of authorship for the Federalist Papers. In the past few decades, the release of a large annotated corpus from the *Wall Street Journal* led to significant increases in the accuracy of statistical parsing (see Marcus, Santorini, and Marcinkiewicz [1993]).

More recently, with the exponential increase in computing power over the past half century and the increased focus on textual methods driven by the requirements of internet search engines, the application of this technique has permeated most disciplines in one way or another. In accounting and finance, the online availability of news articles, earnings conference calls, Securities and Exchange Commission (SEC) filings, and text from social media provide ample fodder for applying the technology.

Can we tease out sentiment from mandated company disclosures and contextualize quantitative data in ways that might predict future valuation components? Can we computationally read news articles and trade before humans can read and assimilate the information? If Twitter's tweets provide the pulse of information, can we monitor these messages in real time to gain an informational edge? Do textual artifacts provide an additional attribute that predicts bankruptcies? Are there subtle cues in managements' earnings conference calls that computers can discern better than analysts? More broadly, can we examine textual artifacts to measure the quantity and quality of information in a collection of text, including both the intended message and, importantly, any unintended revelations? These are all interesting questions potentially answered by the technology of textual analysis.

Textual analysis is an emerging area in accounting and finance and, as a result, the corresponding taxonomies are still somewhat imprecise. Textual analysis can be considered as a subset of what is sometimes labeled qualitative analysis, with textual analysis most frequently falling into the categories of either targeted phrases, sentiment analysis, topic modeling, or measures of document similarity. Readability is another aspect of textual analysis, which is differentiated from some of the prior methods in that it attempts to measure the ability of the reader to decipher the intended message, whereas the other methods typically focus on computationally extracting meaning from a collection of text. Other examples of the more general topic of qualitative analysis would include Coval and Shumway [2001], who consider the information conveyed by noise levels in the Treasury Bond Futures trading pit at the Chicago Board of Trade, or Mayew and Venkatachalam [2012], who examine the audio from earnings conference calls to determine managerial affective states.

Following the pioneering papers by Frazier et al. [1984], Antweiler and Frank [2004], Das and Chen [2007], Tetlock [2007], and Li [2008], accounting and finance researchers have actively examined the impact of qualitative information on equity valuations. The words selected by managers to describe their operations and the language used by media to report on firms and markets have been shown to be correlated with future stock returns, earnings, and even future fraudulent activities of management. Clearly, stock market investors incorporate more than just quantitative data in their valuations, but as the accounting and finance disciplines embrace this new technology we must proceed carefully to assure that what we purport to measure is in fact so.

The burgeoning literature in textual analysis is already summarized well in other papers, although the increasing popularity of the method quickly dates any attempt to distill research on the topic. Li [2010a], in a survey of the literature, provides details on earlier manual-based examples of textual analysis, discusses the modern literature by topical area (e.g., information content, earnings quality, market efficiency), and itemizes a prescient list of potential research topics. His conclusions echo a theme of this paper; that is, the literature needs to be less centered on finding ways to apply off-the-shelf textual methods borrowed from highly evolved technologies in computational linguistics and instead be more motivated by hypotheses “closely tied to economic theories” (p. 158).

Kearney and Liu [2014] provide a more recent survey of methods and literature with a focus on textual sentiment. Their Table 3 provides a useful annotated bibliography of most sentiment-related papers published prior to 2013. Das's [2014] monograph, in addition to reviewing the academic literature, provides an excellent user's guide for someone just approaching the subject, including code snippets for some of the basic tools used in textual analysis.

In what follows we will fold a more selective and focused survey of the accounting, finance, and economics literature on textual analysis into a description of some of its methods. We add value beyond simply offering an updated literature review by also underscoring the methodological tripwires for those approaching this relatively new technique. Qualitative data requires the additional step of translating text into quantitative measures which are then used as inputs into either traditional or text-based methods. We emphasize the importance of exposition and transparency in this transformation process because this is where much of the imprecision of textual analysis is introduced. More generally, we emphasize the importance of replicability in the less structured methods used in textual analysis. Regarding the topic of readability, we underscore the importance of carefully specifying what is meant by the concept in the context of business documents, where the traditional hallmarks of readability (polysyllabic words and long sentences) are rarely distinguishing characteristics in the interpretation of financial text.

The remainder of our survey is organized as follows. In Section 2, before examining those methods intended to extract meaning from text collections, we consider the broader topic of information content and document composition. Regarding information content, we consider how measuring information in qualitative analysis differs from, and is yet in some ways similar to, quantitative measures. Under the topic of document composition, we examine measures of how effectively the textual message is packaged so that its recipient can effectively translate it back into its intended signal. Although document composition includes many items such as formatting and presentation, we focus extensively on readability, an attribute of financial disclosure that is relevant to researchers examining the assimilation of information into asset prices and one that the SEC has historically struggled with.[1] Section 3 of the paper discusses

[1] After many decades of prodding firms to file financial disclosures in a more user friendly form, in 1998 the SEC, under then Chairman Arthur Levitt, created the Plain English initiative. This initiative has been superseded by the broader Plain Writing Act of 2010 which applies to most government documents. See www.sec.gov/plainwriting.shtml.

those methods based on deconstructing a document into a collection of words, where we can then use predefined dictionaries to classify the tone of the documents, train an algorithm using learning methods to identify document characteristics, or determine if there is hidden structure in the form of common topics across a collection of documents. Note that these methods generally ignore the sequence of words in a document. Clearly one of the remaining challenges in textual analysis is to move beyond assuming words occur as independent units, a topic we briefly discuss in Section 4 under the rubric of document narrative. Section 5 considers the measurement of document similarity. In Section 6, we examine issues and limitations in implementing textual analysis, e.g., converting tokens to words, disambiguating sentences, tripwires in parsing specific subsets of financial disclosures, and the issue of levels versus differences. Additionally in this segment, we discuss software alternatives and then outline a simple example that provides someone just approaching this technology with a structure for executing textual methods. Finally, after discussing some promising areas for future research, we provide concluding comments. Our survey is by no means an exhaustive review, but hopefully provides the uninitiated with a gateway into a rapidly evolving branch of our discipline.

2. Information Content, Document Structure, and Readability

Other disciplines with a long history in computational linguistics have rich methodological toolboxes used to assess collections of documents. We will discuss only those methods that, thus far, have had the most impact on the textual analysis literature in accounting and finance.

Before itemizing various methods, we begin with a discussion of the fundamental issue of information extraction in textual analysis and document structure, where the latter is critical in determining whether the consumer of textual data can reasonably extract the information contained in the document. Document structure is reflected by the graphic design of the document along with the writing style used to convey the information. Writing style is usually considered under the rubric of readability—i.e., whether textual information is accessible to the user—and has become a frequently measured attribute of accounting documents. Thus we will discuss the topic of readability in some detail.

2.1 INFORMATION CONTENT

Textual analysis, and more generally qualitative analysis, is most notably demarcated from quantitative analysis by its imprecision. Of course, even in quantitative analysis, we can debate the measurements used to generate data, for example whether earnings should be measured using GAAP or non-GAAP numbers, and the final specification of an empirical model can be but one of many possible permutations. However, quantitative research is sufficiently mature so that there are typically discipline-specific norms for measuring inputs and for selecting methods used to address a given empirical issue.

Traditional quantitative research attempts to identify information contained in a sample within the context of directed hypotheses, where the hypotheses, along with disciplinary traditions, dictate the specific statistical methods and inferential techniques used. In typical quantitative research, while the measure might be imprecise (e.g., does the number associated with Pension Obligations fully reflect the current value of and uncertainty associated with a firm's pension liabilities), the unit of measure is unambiguous.

With textual analysis there is a critical transformation that must take place as we attempt to move from a collection of characters, to extracting the information conveyed by these characters. The meaning of the characters is not unambiguous and in most cases depends substantively on the context of a sentence, document, collection, and when and by whom it was written. Although some areas of natural language state their interpretational objective in terms of the intended message of the document (instructions, consumer information, etc.), keep in mind that our discipline is interested in both the intended and unintended information conveyed by the text. For example, a manager in an earnings call might unintentionally use more weak modal words (e.g., *may*, *could*, and *might*), possibly signaling trouble for the firm.

Consider just a few examples of the potential for imprecision. For those studies that focus on SEC filings, and that wish to use the full extent of the rather limited online time series which begins in 1996, HTML formatting provides a potential source of systemic errors.[2] Document parsing relies on consistency in the structure of the text and any related mark-up language. For many filings prior to roughly 2005, there is a lack of consistency in HTML formatting. For example, Potomac Electric Power in their 20040312 filing use "<TABLE>" tags

[2] Some large firms begin electronic filing in 1994. Electronic filing was required of all firms, with minor exceptions, beginning in 1996.

to define all text paragraphs. A parser that first eliminates tables from a document will fail in accurately parsing this filing. Importantly, the tendency for filings to lack the structural anchors used in parsing is correlated with firm size and time period, thus inducing systematic mismeasurement and not simply noise. In the more general area of qualitative analysis, audio software used to measure stress or other attributes is sensitive to the differing dynamic characteristics of microphones and whether pop filters are used to moderate sibilance.

The magnitude of these errors can be substantial. For example, if *best* is a positive word and the document is not parsed to exclude company names, firms like *Best Buy* will have very positive sentiment measures. If *may* is included as a measure of uncertainty, and the parse does not distinguish between lower and upper case, then you will find an astonishing seasonality in your results. If you use word lists that categorize *mine* and *death* as negative, as some widely used lists do, then you will find the mining and health industries to be extraordinarily pessimistic.

The imprecision of textual analysis is not something that precludes its usage, but is a characteristic that must be confronted in producing empirical results that are expected to have credible impact and that can be reasonably replicated. We will underscore in our discussion that the ideal hypotheses in textual analysis are ones based on straightforward characteristics of the data and requiring the least amount of econometric exorcism to produce the results. We will provide a specific example later in the paper.

In spite of the importance of the initial transformation from text to quantitative summaries, most textual analysis papers in accounting and finance provide vague statements about how a document is parsed and then produce results from a software package where the driving forces behind the results are opaque. Replication of existing studies using textual analysis is, at best, challenging. To assist in replication, we recommend that papers include an appendix providing specific details on the parsing process, an extension facilitated by journals' online appendices.

Although it is impossible to identify the total information content of a collection of words and numbers, the goal of any analysis is to approach this limit. As we extract information from the document text, we must attempt to determine how much context is lost by methods that assume words are independent units, i.e., methods where word sequence is unimportant. And, as we attempt to apply deeper and more complex parsing that does not make this assumption, we

must be careful that the imprecision of the method does not overwhelm any hoped for gains in identifying meaning. As emphasized before, many misclassification errors in textual analysis can do more than add noise to the data, and can unintentionally create latent measures of other firm attributes such as size or industry.

2.2 READABILITY

In this section we consider the overarching issue of whether the receiver of information can accurately reconstruct the intended message. This topic is generally discussed under the general rubric of "readability", but, as we will see, the definition of readability, once we leave the realm of grade-leveling textbooks, becomes elusive.

Related to the topic of readability, is the more general concept of document composition and structure. Given the amount of resources spent on graphic design, presumably the presence of non-textual materials (pictures, graphs, and tables) enhances the ability of the reader to understand the intended message. Certainly studies looking at the impact of financial disclosures could include as a variable of interest the number and characteristics of non-text items. Additionally, the introduction of eXtensible Business Reporting Language (XBRL) in SEC filings and other financial documents provides a structured context for data that will make it easier to computationally interpret the document. Although all of these topics are of interest in terms of document consumption, we will focus specifically on readability as it is the concept most frequently studied in the extant literature.

2.2.1. Examples of Studies using Readability. Before confronting the most critical issue of what is meant by readability, we first discuss prior research on the topic. Research on the readability of accounting narratives has a long history (see Jones and Shoemaker [1994]). Unfortunately, much of the earlier work on readability suffered from small sample sizes or problematic methodologies. For example, Tennyson, Ingram, and Dugan [1990] examine the important relation between financial distress and managerial narrative disclosures. Yet, their analysis focused on comparing text of only 23 US firms who declared bankruptcy with a matched sample of 23 non-bankrupt firms. Lewis, Parker, Pound, and Sutcliffe [1986] provide an analysis of various readability measures (i.e., Fog and Flesch Indexes) for financial reports using only nine Australian firms over a four year period.

The first paper to examine the link between annual report readability and firm performance for a meaningful sample is Li [2008]. In this important and widely-cited paper, Li [2008] measures the readability of annual reports (i.e., Form 10-Ks) using the Fog Index and the number of words contained in the annual report. The Fog Index is a function of two variables: average sentence length (in words) and complex words (defined as the percentage of words with more than two syllables):

$$\text{Fog Index} = 0.4 \text{ (average number of words per sentence + percent of complex words)} \quad (1)$$

As with most traditional readability measures created to differentiate grade school textbooks, the Fog Index equation estimates the number of years of education needed to understand the text on a first reading. Thus, a Fog Index value of 16 implies that the reader needs sixteen years of education—essentially a college degree—to comprehend the text on a first reading.

Li [2008] finds that firms with lower reported earnings tend to have annual reports that are harder to read (i.e., high Fog Index values or high word counts). As noted by Bloomfield [2008], this finding may be caused by poorly performing firms needing to have more text and longer sentences to fully explain their situation to investors. Li also finds that companies with more readable annual reports have higher earnings persistence. The key contribution of Li's paper is linking linguistic features of the annual report to actual firm operating performance.

Following Li [2008], other researchers have used the Fog Index as a measure of annual report readability. Biddle, Hilary, and Verdi [2009] find that firms with high reporting quality (using the Fog Index and two other variables) are associated with greater capital investment efficiency. Guay et al. [2015] find that companies with less readable annual reports (based on six different readability measures including the Fog Index) tend to mitigate this negative readability effect by issuing more managerial forecasts of earnings per share, sales, and cash flows. Miller [2010] finds that small investors trade significantly fewer shares of firms with high Fog Index values and word counts (i.e., less readable annual reports) around the 10-K filing date. Less readable annual reports should be harder to process, especially for less sophisticated investors.

Lawrence [2013] finds that both the Fog Index and the number of words in the annual report are linked to retail investors' stock holdings. His sample includes actual portfolio

holdings for 78,000 US households during 1994-1996. Individual investors are found to invest more in firms whose annual reports contain fewer words and have better readability (as measured by the Fog Index). This result, however, is less pronounced for a group of high frequency trading individuals (more than 48 trades in any year).

Comparing annual reports and earnings press releases written by foreign firms listed on a US exchange with domestic firms, Lundholm, Rogo, and Zhang [2014] find that foreign firms cross-listed in the US produce more readable business documents. They argue that foreign based companies have a greater incentive to make their public documents more readable (i.e., lower Fog Index values) to encourage US investors to invest in their firm. In their Table 2, the summary statistics for the Fog Index are reported separately for foreign and US firms. Interestingly, the mean Fog Index values for earnings press releases are almost identical for foreign firms (16.18) versus US firms (16.24). The authors note that this difference is statistically significant and is consistent, before controlling for other variables, with foreign firms having more readable filings. Yet Fog Index averages that differ by only 0.06 are effectively identical in terms of the number of formal education years needed to understand the text. Similarly, they report that the mean difference in the Fog Index values between the Management Discussion & Analysis (MD&A) section of the 10-K for foreign and US firms is relatively small (17.54 versus 18.06).

Focusing on the link between readability and analyst coverage, Lehavy, Li, and Merkley [2011] find that more readable annual reports, as measured by the Fog Index, have lower analyst dispersion and greater earnings forecast accuracy. They find that 10-K readability is related to how many analysts cover a stock. Firms with higher Fog Index values, after controlling for company characteristics, have more analysts covering the stock. They view this evidence as consistent with "a greater collective effort by analysts for firms with less readable disclosures" (p. 1089). In their reported summary statistics, it is interesting to note that the Fog Index value for the bottom quartile of annual reports is above 18 for each of their sample years, 1995-2006. Generally documents with a Fog Index above 18 are considered unreadable since more than a master's degree is needed to understand the text.

The readability of analyst reports is also associated with investor behavior. De Franco, Hope, Vyas, and Zhou [2015] analyze a sample of over 365,000 annual reports issued during 2002-2009 for readability characteristics. One of their readability measures is an aggregation of

three different readability indexes (Fog, Flesch, and Flesch-Kincaid).[3] The authors find that more readable analyst reports are associated with significantly higher trading volume over a three-day window surrounding the analyst report date.

Some researchers have analyzed business document readability in a laboratory setting. Rennekamp [2012], using 234 participants, varies the readability of the disclosure while holding the length and total information contained in the document constant. Improved readability, based on the SEC's *Plain English Handbook*, is associated with stronger reactions for both good and bad news among the participants. Her study provides an interesting caveat about improving the readability of business documents for retail investors. Rennekamp [2012] finds that "more readable disclosures may cause investors to overreact to information, particularly those who are the least sophisticated" (p. 1322).

From this literature, clearly the role of readability is considered important as a central or adjunct variable in assessing financial documents. The empirical results of these studies repeatedly document a statistical association between a traditional readability measure—i.e., the Fog Index—and other attributes of the firm. In the next section, we question what is meant by readability in the context of financial documents and whether the Fog Index is measuring the intended construct.

2.2.2. Defining and Measuring Readability

The central issue in readability is considering carefully what is meant by the concept in the context of business writing. Although the Fog Index has a long history in grade leveling K-12 textbooks, many have questioned its usage for business documents. Much of the problem centers on how "readability" is defined, which varies across the literature. Jones and Shoemaker [1994] argue that "there is no consensus about how closely the readability measures reflect the actual comprehension process" (p. 172). They also comment that if the words in each sentence were randomly reordered, the passage would be completely unintelligible, yet would have an identical Fog Index value.

[3] All three readability measures are simply linear combinations of sentence length and syllable-related measures. For the latter term, the Fog Index counts "complex words," defined as all words greater than two syllables. Flesch-Kincaid, similar to Fog, produces a grade level measure but uses the average number of syllables per word as the second term. The Flesch Reading Ease score uses the same inputs as Flesch-Kincaid, but scales the linear combination to range approximately from 0 to 100.

Loughran and McDonald [2014] empirically demonstrate that the Fog Index is a poorly specified readability measure when applied to business documents. They argue that in the context of accounting information, definitions of readability focusing on "understanding or comprehension due to the style of writing" (Klare [1963], p. 1) are less appropriate than more general definitions from the literature that focus on "the degree to which a given class of people find certain reading matter compelling and comprehensible" (McLaughlin [1969], p. 639). The simplest and most compelling argument against the use of traditional readability measures in accounting disclosures is the observation that the vast majority of these documents are not distinguished by writing style.

Using a sample of 66,707 annual reports during 1994-2011, Loughran and McDonald [2014] expose a major weakness in the Fog Index. The percentage of complex words (more than two syllables) accounts for half of the Fog Index value. An increase in the percentage of complex words lowers the readability of the document according to the Fog Index. Yet, by far the most frequently occurring "complex" words in business documents are words like *financial, company, operations, management, employees,* and *customers* which are easily understood by investors. They show that syllable counts are a poor measure of readability for business documents. Consistent with this assertion, Loughran and McDonald [2014] find that the Fog Index is not significant in explaining analyst dispersion or earnings surprises.

As a simple proxy for readability of annual reports, Loughran and McDonald suggest using the natural log of gross 10-K file size (complete submission text file) available on the SEC's EDGAR website. This measure is easy to obtain, does not require problematic parsing of 10-Ks, and allows for straightforward replication. They define readability as the ability of investors and analysts to integrate valuation relevant information from the business document into stock prices. Loughran and McDonald find that firms with bigger 10-K file sizes are significantly linked with larger subsequent stock return volatility, analyst dispersion, and absolute earnings surprises. As they note, this association may reflect the underlying complexity of the firm's business. That is, although file size might serve as a proxy for readability, for a given firm it cannot completely separate the fundamental complexity of the firm's business from the language complexity of its annual report.

A number of working papers have used the 10-K complete submission file size as an omnibus measure of annual report readability (see Bonsall and Miller [2014], Bratten, Gleason,

Larocque, and Mills [2014], Ertugrul, Lei, Qiu, and Wan [2015], and Li and Zhao [2014]). Loughran and McDonald note that the simple measure of gross file size correlates highly with more appealing measures such as net file size, where extraneous components have been removed (e.g., HTML or encoded images), or other more specific measures of readability.[4] The gross file size measure also performs at least as well as the alternative readability measures when benchmarked against outcomes such as earnings surprise and analyst dispersion.

Their results underscore that the concept of readability must be delineated in the context of its application. For example, the use of jargon is generally considered a negative attribute in the traditional sense of readability. However, Loughran and McDonald [2014] find that financial jargon positively impacts their measures of readability, thus highlighting that the point of readability in this context is not trying to make financial disclosures readable at the lowest possible grade level.

Alternative readability measures of business communication (Common Words, Financial Terminology, and Vocabulary) are also proposed by Loughran and McDonald [2014]. All of these other metrics outperform the Fog Index in terms of measuring the effective communication of value-relevant information to investors through the annual report. These alternative measures might be useful for studies with a singular focus on document readability, especially for applications to documents in other contexts (e.g., analyst reports or news articles, where file size is less differentiated). Yet, the findings for mandated periodic financial disclosures support using the natural log of gross file size, while controlling for firm size, as a simple, but effective proxy for readability. As they note, it is impossible to entirely disentangle readability from complexity, so these proxies should be interpreted within this broader frame. Paralleling this conclusion, You and Zhang [2009] find that longer annual reports have a delayed investor reaction over the 12 months following the filing date. Form 10-Ks with higher annual report word counts appear to reduce the ability of investors to quickly incorporate information into current stock prices.

[4] At first glance a reader would assume that file size net of HTML markup, binary segments, etc. would be more appropriate. Along those lines, Bonsall et al. [2015] argue that log file size has significant measurement error as a proxy for readability. However, Loughran and McDonald [2014] note that net file size is highly correlated with gross file size and has little impact on their results. Thus they opt for gross file size which avoids any subjective parsing rules and is easily available.

Clearly log file size is not a perfect measure of 10-K readability. For example, there is evidence in the literature that firms responded to the Enron accounting scandal by expanding the number of pages in their annual report to improve their firm-specific transparency. Leuz and Schrand [2009] find that this increase in document length actually lowered the firm's cost of capital.

As with all text measures, the use of readability measures must consider the context of application. Although file size proxies make sense for documents such as 10-K filings, it is less clear that this is a distinguishing feature of the text from an earnings conference call. In cases like this, where the length of the text is not highly variable, some of the other approaches documented in Loughran and McDonald [2014] that focus more on content are likely more appropriate.

In sum, researchers using readability as a measure must be careful to identify the variable's intent. If it is simply intended to be an omnibus measure capturing the overall complexity of the firm, then measures such as the log of gross file size, Common Words, or Vocabulary are reasonable proxies. If, instead, the intent of the variable is to specifically measure the reader's ability to assimilate the document's written message, then the researcher must carefully define what is meant by this concept. Is a good financial document one that is readable by someone with a lower grade level or is it one targeted toward analysts and rich with jargon and data? Specificity of this intent will, to a significant degree, dictate the characteristics of an appropriate measure.

The use of the widely popularized Fog Index is likely inappropriate. Although this variable is in some cases correlated with underlying business attributes, Loughran and McDonald [2014] emphasize that the Fog Index is difficult to correctly measure in business documents and can be misspecified. The two components in the Fog Index, complex words and sentence length, are negatively correlated in their sample, which is not consistent with their presumed effects. In business documents, complex words are typically not challenging words and are more likely to signal a specific industry (e.g., pharma). Additionally, the parsing of business documents into sentences is error prone—a problem we will discuss in a subsequent section.

Importantly, if the intention is to measure document readability, then researchers face the difficult identification problem of separating the business and the document. These issues are intertwined because the document attempts to describe the economic reality of the business. As

argued by Leuz and Wysocki [2016], this is a fundamental problem that plagues all accounting quality metrics, be it accruals, earnings management, or text based measures. All measures of readability are subject to this problem. Perhaps we would be better off focusing on the broader topic of information complexity and avoiding the term readability, which is constrained by its historical context.

3. Bag of Words Methods and the Term-Document Matrix

While readability focuses on the individual's ability to comprehend a message, the methods we now focus on attempt to computationally distill meaning from the message. Interestingly, given the inherent nature of language and writing, some of the most widely used textual methods rely on the critical assumption of independence to reduce the extraordinary dimensionality of a document, where independence means that we assume the order, and thus direct context, of a word is unimportant. Methods where word sequence is ignored are typically labeled as "bag of words" techniques. Many of these are based on collapsing a document down to a term-document matrix consisting of rows of words and columns of word counts. Given the vast methodological tool chest of computational linguistics, or the option of morphing traditional econometrics into the qualitative realm, tabulating word counts would seem to be a baby step in the science of applying textual analysis.

To the extent that the sequence of words in a document is not critically important to the attribute of interest, the use of word counts allows the computational task of summarizing a large document to be simplified by orders of magnitude. The critical question for methods going forward is whether important incremental information can be extracted by more deeply parsing for contextual meaning. This is essentially a signal to noise tradeoff, where the nuance of context is the signal and the increasing imprecision of deep parsing is the noise.

3.1. TARGETED PHRASES

One of the simplest, but at the same time the most powerful, approaches to textual analysis is facilitated by hypotheses that allow the researcher to target a few specific words or phrases. Because of ambiguity, large word lists are much more prone to error when compared to tests focusing on a few unambiguous words or phrases. For example, Loughran, McDonald, and Yun [2009] consider the frequency of the word "ethic" (and its variants) along with the phrases

"corporate responsibility", "social responsibility", and "socially responsible" in 10-K filings to determine if these counts are associated with "sin" stocks, corporate governance measures, and class action lawsuits. They find that firms whose managers are more focused on discussing these topics are firms more likely to: be labeled as sin stocks, have low corporate governance measures, and be sued in the year subsequent to the filing. The parsing required to achieve results such as these is relatively straightforward and easily replicated. In a subsequent section, we provide an example of textual analysis using this simple context where we look for the term "non-GAAP" in financial disclosures. Unfortunately, many of the interesting hypotheses relating to textual analysis do not provide such a sharp and measureable hypothesis.

3.2 WORD LISTS

The next evolutionary step beyond simple targeted phrases is compiling word lists that share common sentiments (e.g., positive, negative, uncertain). Armed with such lists, a researcher can count words associated with each attribute and provide a comparative measure of sentiment. Of course the challenge of this technique arises as a result of homographs (words with different meaning, but the same spelling) and context.

Technically, a "dictionary" is a tabulated collection of items, each with an associated attribute, as, for example, in its traditional form of a word and associated definition. Thus our discussion should be restricted to the term "word lists" where we are simply creating collections of words that attempt to identify a particular attribute of a document. For our purposes, the distinction is not critical and we will use the two terms interchangeably. Also note, that in much of the literature, dictionaries created for very specific purposes (versus a generic list of words), are often referred to as lexicons. The dictionary methodology is another example of the "bag of words" approach, underscoring that the tabulation of words from a document into lists discards all information that can be distilled from word sequences.

In measuring the tone or sentiment of a financial document, researchers typically count the number of words associated with a particular sentiment word list scaled by the total number of words in the document. Thus, for example, higher proportions of negative words in a document indicate a more pessimistic tone. For researchers, the first step in the process is to decide which dictionary should be employed to tabulate the proportion of targeted attributes. For example, the Harvard General Inquirer word lists, a group of lists used historically in the

sociology and psychology literature, purports to measure more than 100 attributes of a document including pleasure, pain, arousal, overstated, political, interpersonal relations, and need.[5]

The use of dictionaries to measure tone has several important advantages. First, once the dictionary is selected, researcher subjectivity is avoided. Second, since computer programs tabulate the frequency counts of words, the method scales to large samples. Third, with publicly available dictionaries, it is more straightforward to replicate the analysis of other researchers. As we will discuss in a later section, an important component of classifying words is identifying the most frequently occurring words within each classification—i.e., those words most influential in the final tally.

In the accounting and finance literature, four different word lists have been extensively used by researchers: Henry [2008], Harvard's General Inquirer (GI), Diction, and Loughran and McDonald [2011].[6] While researchers primarily focus on positive and negative word lists, the dictionaries also generally include targeted subcategories of word themes like uncertainty, weak modal, constraints, pleasure, pain, extreme emotion, and even virtue. Although both the Diction and Harvard GI word lists were not created with financial text in mind, researchers have used the lists to measure tone in newspaper articles, earnings conference calls, annual reports (Form 10-Ks), IPO prospectuses, and press releases.

3.2.1. The Henry [2008] Word List. The first word list we are aware of that was created for financial text specifically is Henry [2008]. The positive aspect of the Henry [2008] word lists is that her dictionaries were created by examining earnings press releases for the telecommunications and computer services industries. The obvious weakness of her list is the limited number of words contained in the list. For example, the Henry list has only 85 negative words while the Harvard word list contains more than 4,100. Commonly occurring negative words in business communication like *loss*, *losses*, *adverse*, and *impairment* are surprisingly missing from her list. Managers have many more ways to imply negative tone in business communication than the 85 words on Henry's negative list.

[5] See http://www.wjh.harvard.edu/~inquirer/homecat.htm.

[6] These four dictionaries are obviously not a complete list. For example, Larcker and Zakolyukina [2012] create self-contained word categories to gauge deceptive language by managers during earnings conference calls. They have subcategories measuring hesitations (*hmmm*, *huh*, and *umm*), extreme negative emotions (*idiot*, *slimy*, and *disgraceful*), and extreme positive emotion (*tremendous*, *smashing*, and *swell*). Matsumoto, Pronk, and Roelofsen [2011] create a list of financially oriented words. Bodnaruk, Loughran, and McDonald [2015] produce a list of 184 financially constraining words.

Price, Doran, Peterson, and Bliss [2012] use the Henry [2008] word lists to gauge tone during quarterly earnings conference calls for publicly-traded stocks. They report that during both three-day and two-month windows, firms with positive tone in the question-and-answer portion of the conference call experience significantly higher stock returns. Conversely, conference calls with negative tone, as measured by the Henry [2008] lists, have negative abnormal returns. Price et al. [2012] assert that the Henry [2008] dictionaries are better at measuring the tone of earnings conference calls than the Harvard IV-4 word lists since only the Henry lists document a significant market reaction in both the initial reaction window and in the 60-day drift period. Similarly, Doran, Peterson, and Price [2012] use the Henry [2008] word lists to focus on the earnings conference calls of Real Estate Investment Trusts. They find that the tone of the conference call is significantly linked with stock returns during the conference call even after controlling for the earnings surprise.

Davis, Ge, Matsumoto, and Zhang [2015] examine manager-specific optimism during earning conference calls. They use the Henry [2008], Diction, and Loughran and McDonald [2011] positive and negative word lists to gauge sentiment. The authors report that manager-specific tone is positively linked with future operating performance using the Henry [2008] and Loughran and McDonald [2011] word lists. Tone generated from the Diction word lists is not associated with subsequent return-on-asset values.

3.2.2. Harvard General Inquirer Word Lists. Initially in the disciplines of accounting and finance, most researchers used the Harvard GI and Diction word lists for the simple reason that these lists were the first ones readily available. Accounting for inflections (i.e., different forms of the same word), the Harvard negative word list contains 4,187 words. In a highly influential paper, Tetlock [2007] links the tone of the *Wall Street Journal's* "Abreast of the Market" daily column with stock market levels. He finds that high levels of journalistic pessimism in the daily column are related to both lower subsequent stock returns and to higher subsequent stock market volatility. Interestingly, the downward pressure on stock prices is not caused by the *Wall Street Journal* column providing new fundamental information on company valuations.

Instead, Tetlock [2007] proposes that the "Abreast of the Market" column proxies for investor sentiment. Higher investor pessimism temporarily lowers the level of the Dow Jones Industrial Average (Dow). Like most papers in the textual analysis literature examining stock

returns, the economic magnitude of the soft information is somewhat limited. Tetlock [2007] finds that a one-standard deviation increase in pessimism is related to only an 8.1 basis point decline in the Dow the following day.

Following Tetlock [2007], a number of papers use the Harvard IV-4 positive and negative word lists to gauge tone of newspaper columns. For example, Tetlock, Saar-Tsechansky, and MacSkassy [2008] examine *Wall Street Journal* and *Dow Jones News Service* stories on S&P 500 firms. They find that a higher frequency of negative words in firm-specific news stories is linked with lower subsequent earnings even after controlling for trailing accounting information and Wall Street analyst forecasts. Using a sample of more than 900,000 Thomson-Reuters news articles, Heston and Sinha [2015] find that a positive Harvard net sentiment measure (positive word frequencies minus negative word frequencies) for an article mentioning a specific company produces high returns for the same firm one to two days after the article's publication. Conversely, they find that firms with negative sentiment news stories are associated with lower short-term stock returns.

Utilizing the Harvard IV-4 negative and positive word categories, Kothari, Li, and Short [2009] examine the content of disclosures by firms, analysts, and news outlets. They find that disclosure tone is associated with both stock return volatility and analyst forecast error dispersion. More positive disclosures by the firm, analysts, or the media are linked with lower volatility and forecast dispersion. In contrast, negative new information contained in disclosures is associated with significantly higher volatility and analyst dispersion.

Using a large sample of initial public offerings during 1996-2005, Hanley and Hoberg [2010] examine how the tone of the initial prospectus (Form S-1) impacts pricing and first-day returns. The authors gauge prospectus tone using the Harvard IV-4 positive and negative word lists. In the Risk Factors section of the prospectus, Hanley and Hoberg [2010] find that more positive net tone (% positive minus % negative) is linked with lower first-day returns and smaller changes in the offer price revision. They argue that investors view positive tone written by managers and underwriters, which face legal penalties for misstatements, as a credible signal concerning the riskiness of the offering.

3.2.3. Diction Optimism and Pessimism Word Lists. Like the various Harvard GI word categories, Diction has 35 different dictionary subcategories.[7] To create a positive word list, researchers typically combine the Diction optimism subcategories of praise, satisfaction, and inspiration. For negative words, the Diction pessimism subgroups of blame, hardship, and denial are pooled together. Using this approach, there are 686 Diction optimism words and 920 Diction pessimism words. In measuring document tone, accounting researchers have been much more active users of the Diction word lists than finance researchers.

Davis, Piger, and Sedor [2012] find that firms with more positive tone (using the Diction word lists) in their earnings press releases are associated with higher subsequent return on assets (ROA). Their paper proposes that the language managers use to describe operations in earnings press releases provides a direct but subtle signal about management's expectations of their future performance. The more positive the tone (i.e., % of Diction optimism words minus % of Diction pessimism words) of the earnings press release, the higher is the firm's ROA in the four subsequent quarters. Similarly, Davis and Tama-Sweet [2012] find a significant linkage between tone in the MD&A section of the Form 10-K and future ROA. The more pessimistic the MD&A tone, the lower is subsequent ROA for the company.

Instead of linking earnings announcement tone with subsequent operating performance, Rogers, Van Buskirk, and Zechman [2011] examine the relation between Diction net tone and shareholder litigation. Using a matched pair methodology, they find that companies with higher optimism in their earnings announcements are associated with significantly higher probabilities of being sued by their shareholders. It is reported that a one-standard deviation increase in net Diction optimism is related to a 52% increase in the likelihood of being sued by shareholders.

3.2.4. Limitations of the Harvard and Diction Sentiment Word Lists. Li [2010b] and Loughran and McDonald [2011] criticize the use of Harvard IV-4 and Diction lists to gauge managerial tone in corporate filings. For example, Li [2010b] finds no positive relation between the tone of the MD&A section of the 10-K (i.e., annual report) using the GI and Diction dictionaries and future performance. Separately, Loughran and McDonald [2011] report that almost 75% of the Harvard GI negative words do not have pessimistic meaning when used in the context of financial documents.

[7] The Diction word lists are available for purchase from www.dictionsoftware.com.

Loughran and McDonald [2011] argue that Harvard IV-4 negative words like *tax, cost, capital, board, liability,* and *depreciation,* which are predominate in firms' 10-K filings, do not typically have negative meaning when appearing in an annual report. They also document that several of the Harvard negative words are likely to proxy for specific industries. For example, management's use of *crude*, *cancer*, and *mine* do not have negative meaning and merely proxy for the oil, pharmaceutical, and mining industries. They warn that researchers attempting to measure sentiment in business communications should not use "classification schemes derived outside the domain of business usage" (p. 62). Instead, word lists designed specifically for business communication should be used to measure the sentiment of business text.

Analyzing the Diction optimistic and pessimistic words, Loughran and McDonald [2015] likewise find that the vast majority of the Diction words are likely misclassified. Frequently occurring Diction optimistic words like *respect, necessary, power,* and *trust* will not typically have positive meaning when used by managers to describe future or current operations. The two authors also question whether Diction pessimism words like *no, not, without, gross,* and *pain* will have negative meaning in the context of the typical accounting disclosure.

3.2.5. Loughran and McDonald [2011] Word Lists. Loughran and McDonald [2011] created six different word lists (negative, positive, uncertainty, litigious, strong modal, and weak modal) by examining word usage in a large sample of 10-Ks during 1994-2008.[8] Their approach was to "create a relatively exhaustive list of words that makes avoidance much more challenging" (p. 44). They create the sentiment lists based on the most likely interpretation of a word in a business context. The Loughran and McDonald (LM) word lists are quite extensive: their dictionaries contain 354 positive and 2,329 negative words. The LM lists haves two main advantages over the other three word dictionaries commonly used in the accounting and finance literature. First, unlike the Henry [2008] list, they are relatively comprehensive. Generally, no commonly appearing negative or positive words are missing. Second, the LM lists were created with financial communication in mind. The only words that potentially could enter their dictionary are ones actually used by managers in 10-Ks.

As noted in the Kearney and Liu [2014] textual analysis review paper, "the L&M lists have become predominant in more recent studies" (p. 175). Typically, papers have used the LM

[8] The Loughran and McDonald word lists are available at http://www.nd.edu/~mcdonald/Word_Lists.html.

word lists (primarily negative words) to gauge the tone of the business communication. For example, Feldman, Govindaraj, Livnat, and Segal [2010] use the LM positive and negative word lists to examine the market's immediate response to changes in MD&A tone for a large sample of 10-K and 10-Q filings. The authors find higher stock market returns when changes in tone are more positive even after controlling for earnings surprises and accruals.

Many papers have used the LM word lists to measure tone in newspaper articles/columns. Expanding on the earlier work of Tetlock [2007], Dougal, Engelberg, Garcia, and Parsons [2012] examine the authorship of the *Wall Street Journal's* (*WSJ)* "Abreast of the Market" column. They find that journalists associated with more pessimistic column tone are directly linked to more negative market returns the following day.

Examining the influence of the media on 636 acquisitions that had a negative announcement reaction from investors during 1990-2010, Liu and McConnell [2013] find that managers are sensitive to their reputational capital. They report that both the level of media attention (i.e., number of articles) on the proposed acquisition and the tone of the corresponding news articles (using percentage of LM negative words) is significantly linked with the probability of abandoning the deal.

Garcia [2013] uses both the LM positive and negative word lists to measure the tone of two financial columns in the *New York Times* during 1905-2005. He finds that newspaper sentiment plays a role in predicting future stock returns particularly during recessionary periods. Using the frequency of LM negative words to quantify tone in newspaper articles, Gurun and Butler [2012] document a link between local advertising dollars and local media slant. Local newspapers use significantly fewer negative words in articles about local firms than stories on non-local companies.

Sentiment analysis on newspaper articles can also uncover the media's role in investors' mistaken tendencies to chase mutual funds with high past returns. Solomon, Soltes, and Sosyura [2014] find that investors chase funds with high past returns only if the funds received media coverage on their holdings. The authors also find that fund-specific newspaper articles with more positive tone (using LM positive and negative word lists) are linked with higher quarterly investor capital inflows for those funds.

Instead of using newspaper articles for investment advice, millions of investors look to user generated opinions from Seeking Alpha (SA).[9] Chen, De, Hu, and Hwang [2014] find that the tone (using LM negative word frequencies) of opinions contained in SA commentary are associated with future stock returns and even subsequent earnings surprises. Focusing on corporate press releases, Solomon [2012] examines the role investor relations (IR) firms play in their client's media coverage. Consistent with a "spin" hypothesis, he finds that IR firms enhance media coverage for good news press releases relative to negative news releases (using the LM negative word list). Can companies strategically increase their stock prices by using press releases prior to a merger announcement? Ahern and Sosyura [2014], in a sample of 507 acquisitions during 2000-2008, find that fixed exchange ratio bidders attempt to bump up their stock prices during the private negotiation phase of the merger. Company press releases by bidders lead to increased media coverage with a more positive tone (using the LM positive/negative word lists) and a slight increase in the acquirer's stock price.

Whether managers attempt to inform or mislead investors by their language usage in earnings press releases is an open question. Huang, Teoh, and Zhang [2014] find strong evidence that the tone of an earnings press release actually misinforms market participants. Using a large sample during 1997-2007, they report that abnormal positive tone, using LM positive and negative words, in the earnings press release is significantly linked with poor subsequent earnings and cash flows for up to three years after the initial release.

Does the content and pitch of a manager's voice during their discussion with analysts in a conference call provide any insights for the company's contemporaneous returns or even future performance? Using a sample of earnings conference call audio files during 2007, Mayew and Venkatachalam [2012] measure positive and negative aspects of managers' emotional state. In regressions with contemporaneous stock returns as the dependent variable, they find that LM positive word frequencies are associated with higher returns while LM negative word frequencies are linked with lower returns. Importantly, the stock market responds to the vocal cues of managers during the Q&A portion of the call even after controlling for the tone of the call. Positive managerial affect is associated with higher contemporaneous returns.

The Loughran and McDonald word lists have also been used to gauge the tone of mutual fund letters to shareholders (Hillert, Niessen-Ruenzi, and Ruenzi [2014]), IPO prospectuses

[9] www.seekingalpha.com is a website providing financial analysis and news about the financial markets.

(Ferris, Hao, and Liao [2013] and Loughran and McDonald [2013]), and analyst reports (Twedt and Rees [2012]). A promising technique to adapt existing dictionaries to alternative media is contained in Allee and DeAngelis [2015]. The authors use the LM [2011] positive and negative word lists to measure tone in conference calls. However, they fine tune the LM word lists to the conference call setting by removing "question" as a negative word. The authors also do not count tokens like "good" as a positive word if it is followed by "morning", "afternoon", "day", or "evening" and drop "effective" if it comes before "income", "tax", or "rate".

3.2.6. Zipf's Law. The driving force behind a critical tripwire in word classification is that word counts tend to follow a power law distribution, a phenomenon frequently referred to as Zipf's law (see, for example, section 1.4.3 of Manning and Schütze [2003]). That is, given the power law nature of word count distributions, certain words can potentially have a large impact on the results.

For illustration, we plot the relative frequency (word counts / total words) for all 10-K/Q type SEC filings from 1994-2012 for all words and then for the subset of Loughran-McDonald negative words in figure 1. Clearly, the word counts for all words and any non-pathological subset of words is dominated by the top few entries. The dominant words for the all-inclusive group are typically labeled stop words. For example, the first five words in this group are *the, of, and, to,* and *in*. Given that most business applications of textual analysis focus on using word counts from sentiment categories, the elimination or special treatment of stop words is typically not necessary.

Consistent with the pattern in figure 1, the cumulative percentage of the top 25 negative words occurring in 10-K/Q filings relative to the total count of all 2,329 negative words is about 44%, i.e., about 1% of the negative words account for about 44% of the negative count. That very few words will dominate the tabulation of any count of categorical words is a critical characteristic of this method. Research using word classifications must identify the proportions of the most frequently occurring words so that the reader can determine if misclassification is driving the paper's results. Although to some extent this determination is subjective, at a minimum this approach allows readers to determine if any patently misclassified words are driving the results. For example, as reported in Loughran and McDonald [2011] using the Harvard dictionary, if purportedly negative words like *vice* (president), *board, liability, tire,* and

depreciation are among the most commonly occurring negative words, it is unclear that the researcher is really measuring pessimistic tone.

3.2.7. Term Weighting. In applications using vectors of word counts (bag of words), a substantial literature in computational linguistics debates how these counts should be normalized (see, for example, Salton and Buckley [1988] or Zobel and Moffat [1998]). In most instances we do not want to use the raw count, since this is obviously strongly tied to document length. A simple use of proportions solves this problem, but in some instances we might also want to adjust a word's weight in the analysis based on how unusual the term is. For example, among LM negative words, the word *unfavorable* appears 1,000 times more often than *expropriating*, *misinform*, or *indict* in the periodic disclosures of firms. Perhaps the more unusual words should receive more weight in the tabulation of negative sentiment.

Loughran and McDonald [2011] consider one of the more common term weighting schemes from the literature labeled *tf-idf* (term frequency-inverse document frequency). Define df_t as the number of documents in a collection of documents containing the term *t*. Let *N* represent the total number of documents in the collection. The inverse document frequency is then:

$$idf_t = \log \frac{N}{df_t} \tag{2}$$

If $tf_{t,d}$ is the raw count of term *t* in document *d*, and a_d is the average word count in document *d*, then:

$$tf\text{-}idf_{t,d} = \begin{cases} \frac{\left(1 + \log\left(tf_{t,d}\right)\right)}{(1 + \log(a_d))} \; log \frac{N}{df_t} & if \; tf_{t,d} \geq 1 \\ 0 & otherwise \end{cases} \tag{3}$$

Although Loughran and McDonald [2011] find that this approach produces regressions with better fit than those using simple proportions, and other papers such as Brown and Tucker [2011] have also applied the transformation, most papers have not considered this modification. Given there are many possible weighting schemes available in the existing computational linguistics literature, this aspect of the method allows the researcher too many degrees of freedom in pre-

selecting the final empirical model. The potential to increase the power of these tests, however, suggests that such methods should be carefully studied.

To summarize our discussion of word lists, many studies have relied on sentiment classification dictionaries derived in other disciplines as the textual analysis literature has evolved in accounting and finance. The results of Loughran and McDonald [2011] indicate that such applications can produce spurious results. In many studies, such as Loughran and McDonald [2011 and 2015] or Twedt and Rees [2012], the empirical results using the different dictionaries are often very similar. However, the use of word lists derived outside of the context of business applications have the potential for errors which are not simply noise and can serve as unintended measures of industry, firm, or time period. The computational linguistics literature has long emphasized the importance of developing categorization procedures in the context of the problem being studied (e.g., Berelson [1952]).

At the same time, applying the Loughran and McDonald dictionaries, which were derived in the context of 10-K filings, without modification to other media such as earnings calls and social media is likely to be problematic. Using the Loughran and McDonald dictionaries as a common base with explicit modifications based on medium is a solution that can avoid some of the issues associated with sentiment measures. Additionally, identifying and reporting words that dominate the word counts can reduce the likelihood of misclassification. And, more carefully considering how the terms are weighted in the sentiment counts could improve the power of statistical tests attempting to identify sentiment patterns.

3.3 NAÏVE BAYES METHODS

Among alternative approaches for word classification using supervised machine learning, such as N-grams and support vector machines, most popular would be the Naïve Bayes method. Naïve Bayes has several purported advantages. First, it is one of the oldest, most established methodologies to analyze text. Second, because machines, instead of humans, read the text for content, large corpuses of data can easily be included in the analysis. Third, once the rules/filters of gauging the text are established, no additional researcher subjectivity affects the measuring of tone in the business communication document.

The main weaknesses of the Naïve Bayesian methodology is the difficulty of others to replicate the results. Since the Naïve Bayes procedure has literally hundreds if not thousands of

various unpublished rules/filters to measure the context of documents, other researchers would be challenged to replicate the results. Typically, a limited number of sample documents are analyzed to teach the program, for example, which sentences in an annual report are classified as being "negative," "positive," or "neutral." The initial sample of documents used to generate the classifications frequently is not provided.

The earliest use of the Naïve Bayes approach in finance is Antweiler and Frank [2004]. They examine 1.5 million stock message postings on Yahoo! Finance and Raging Bull for a small number of firms. Initially, a sample of only 1,000 internet stock message postings was used to train the filters for the program. Although message postings have only a limited impact on stock returns, Antweiler and Frank [2004] find that the number of posted messages is linked with subsequent stock return volatility. As might be expected, higher disagreements among the postings are associated with higher subsequent trading volume. Likewise, Das and Chen [2007] use textual analysis to measure sentiment in message board postings for 24 high-tech stocks. They find that stock message board postings are related to stock market levels, trading volume, and volatility.

Li [2010b] uses the Naïve Bayes method to examine the content of forward-looking statements (FLS) in the MD&A section of the 10-K. Initially, 30,000 randomly selected sentences are manually coded by business students at the University of Michigan and then the sentences are used to train the Naïve Bayes learning algorithm. He finds that the average tone of the FLS is positively linked with subsequent earnings. More positive tone when discussing future operations is associated with higher future earnings for the firm.

Under a Naïve Bayesian framework, Jegadeesh and Wu [2013] determine the relative word weights on the basis of the market's reaction to the 10-K filing. Importantly, their creative approach uses the market's reaction to the 10-K filing to determine the classification (i.e., positive or negative) of each word. Thus, their methodology removes researcher subjectivity from the decision of which words to include in the analysis. Unlike much of the prior literature, Jegadeesh and Wu [2013] find that positive tone is related to the market's reaction of the annual report filing.

Similarly, Purda and Skillicorn [2015] data mine 10-Ks to find words that are best at predicting fraud. Although expected words like *acquisition* and *acquisitions* appear high on the list, they also report many instances where the most predictive words (e.g., *at, as, it, or, on, may*)

might fail “smell tests” if researchers were to create an ex ante list of fraud markers. The approach of Purda and Skillicorn relies heavily on deviations of the firm from itself, highlighting the role (and potential potency) of using the company as its own control.

Huang, Zang, and Zheng [2014] use the Naïve Bayes machine learning approach to gauge the sentiment contained in 363,952 analyst reports. Their trained Bayes algorithm categorizes more than 27 million sentences from analyst reports into three categories: positive, negative, and neutral. A handful of additional positive sentences in the analyst report are associated with a large and significant impact on a firm’s earnings growth rate five years after the publication of the report.

Using the Naïve Bayes algorithm, Buehlmaier and Whited [2014] model the probability of a firm being financially constrained on the basis of the MD&A text in a 10-K. They find that more financially constrained firms are associated with higher stock returns. Surprisingly, they find that the largest, most liquid companies are the ones most affected by financial constraint risk. Buehlmaier and Zechner [2013] use the Naïve Bayes methodology to measure sentiment in newspaper articles concerning US merger announcements. They show that the information about sentiment contained in the news media stories only slowly dissipates into stock market valuations.

The nature of this method makes important the need for the researcher to fully reveal the words driving the empirical classifications. Such disclosure would allow other researchers to determine if the results are possibly keying off of an unintended word or phrase which acts as a flag for a particular industry or time period.

Note that methods such as Naïve Bayes or the inverse regressions of Taddy [2015] can be viewed as another means of identifying and weighting sentiment words. Whether the simple algorithmic transformation of sentiment counts such as tf-idf or the approach of using statistical methods to identify and estimate the weights for words produces the most discerning classification of sentiment is yet to be determined.

3.4. THEMATIC STRUCTURE IN DOCUMENTS

Still within the “bag of words” realm are techniques that can be used to classify common themes in documents or simply identify themes within a corpus of documents. Broadly, these techniques, like most, are attempting to reduce the dimensionality of the term-document matrix,

in this case based on each word's relation to latent variables. In simple terms, we can think of these techniques as essentially factor analysis for words. The evolution of these techniques has been accelerated by their usefulness in search engines.

One of the earliest approaches to this type of classification is latent semantic analysis (LSA)—also known as latent semantic indexing—where the term-document matrix is reduced using singular value decomposition. For internet search firms, such as Google, this technique is useful because it is able to see one page discussing automobiles as similar to another page discussing cars, while at the same time rejecting a page discussing cumulative abnormal returns (CARs) based on how the words load on latent variables. To our knowledge, LSA was first used in business by Boukus and Rosenberg [2006], who analyze the information content of the Federal Open Market Committee's minutes. The distinguishing feature of this method is that it can avoid the limitations of count-based methods associated with synonyms and polysemy (terms with multiple meanings).

The concept of LSA has evolved, first with its extension to probabilistic latent semantic analysis (pLSA) based on a latent class model (see Hofmann [2001]) and then with Dirichlet based priors in latent Dirichlet allocation (LDA, see Blei, Ng, and Jordan [2003]).[10] LDA allows the researcher to identify latent thematic structure within a collection of documents. LSA and LDA have in common the use of the term-document matrix, reducing the dimensionality of the term space to a user specified magnitude, and producing concept or topic weights. They differ in their estimation framework. While LSA uses singular value decomposition to identify an orthogonal basis within the dimensionality constraint, LDA uses a Bayesian model that views the documents as a mixture of latent topics (see Crossno et al. [2011] for a more detailed comparison). Essentially LDA is a generative model that identifies a topic model, which best represents the data. A constraint of both the LSA and LDA techniques is that they work best when applied to large documents

Huang et al. [2015] provide one of the first applications of this method in accounting and finance, using the technique to examine the topical differences between conference call content and subsequent analyst reports. Whereas the traditional use of announcement returns made it difficult to separate out the amount of incremental information actually provided by the analysts,

[10] A Dirichlet prior is essentially a multi-variate version of a beta distribution. pLSA can be shown to be equal to LDA under a uniform Dirichlet prior.

by comparing topical differences the authors are able to isolate the value added of analyst reports. They document that analysts provide significant and differentiated information beyond that contained in the conference call. The level of detail they provide in documenting this technique and the factors driving the results serves as a canonical example of how new textual methods should be introduced into the literature.

4. Document Narrative

In the methods we have discussed thus far, the assumption that interpretation is independent of word sequence substantially simplifies and expands the methods that can be applied. Presumably professional writers would argue that although word choice is important, the essential character of any text is based on how the story is told through the sequencing of words. Progress toward the ideal of deeply parsing for meaning is faced with a tradeoff. The benefit is the value added as a result of meaning derived from word context and grammatical structure. This benefit is offset by the cost of imprecision associated with attempts to computationally derive meaning.

To better understand the progression from collections of symbols to knowledge, we would suggest the following hierarchy of analysis: lexical, collocation, syntactic, semantic, pragmatic, and discourse. Take, for example, a sentence from Google's 2014 10-K: "That leaves out billions of people." (p. 4). The first step in analyzing text, *lexical*, is parsing the document's characters into chunks of words or meaningful tokens, which is straightforward in this simple example. We discuss in a subsequent section, for more typical examples involving a body of text, how this step requires care when deciding which tokens are considered words.

Next in the hierarchy is *collocation*. For some words, much of their meaning is derived from their collocation with other words. The bigram of "going" and "concern" is an example where collocation is important, and if we extend this to n-grams, when this bigram is preceded by "substantial", "doubt", "ability", and "continue", the phrase is then a principle statement in accounting. In the Google example, the bigram "leaves out" is common relative to a random pair of words. The phrase "leaves out", according to Google's Ngram Viewer, occurs with a relative frequency of $4x10^{-5}$ percent versus a meaningless phrase like "far next" which occurs only $5x10^{-8}$ percent of the time (see https://books.google.com/ngrams).

If we can identify a collection of words as a sentence, then using *syntactic* analysis we can derive additional information by examining the grammatical structure of the sentence. In the Google example, syntactic analysis would allow us to determine that "billions" refers to "people".

Beyond syntax, *semantics* attempts to infer meaning within the context of the sentence. In the Google example, we would only understand what the basic meaning of the sentence was and could determine, for example, that "leaves out" was not discussing the removal of tree waste from gutters. *Pragmatics* infers meaning from information immediately preceding and following the sentence, in addition to context provided by external knowledge. Pragmatics, in the Google example, would allow us to understand from the context of the paragraph and the reader's broader understanding of online activity, that "that" refers to the large number of people without access to the internet.

And, finally, *discourse* is the attempt to derive meaning from the collective document. In the Google case, we could then infer that the company is trying to highlight potential growth opportunities based on the expansion in internet users. If potential growth was being highlighted because of market concerns about this valuation attribute going forward, such a conclusion would be an even deeper example of meaning derived from the text.

Thus far, applications in accounting and finance are predominately in the initial phase of this interpretive sequence of lexical to discourse analysis. One of the fundamental challenges of artificial intelligence is to traverse this sequence. Clearly this broader attempt at extracting meaning from documents still has far to go in achieving such a higher level of comprehension.

Allee and DeAngelis [2015] provide a good example in accounting and finance of a first step beyond the bag of words approach by measuring tone dispersion, an approach they label as "a parsimonious measure of narrative structure" (p. 242). Tone dispersion is the extent to which tone is concentrated or spread across a document. High tone dispersion suggests that good or bad news is pervasive, while low dispersion would suggest a more isolated issue. They argue that more dispersed tone should amplify the impact of either good or bad news. Using earnings conference calls, they find that tone dispersion is related to firm performance, financial reporting choices, and the incentive to manage firm perception.

Another example of moving beyond word counts is provided by Chen and Li [2013], who consider the role of estimation in the accruals portion of earnings as reflected in the notes to

financial statements. They classify three types of linguistic relations centering on the root word "estimate" and its variants, and then use Stanford's open source statistical parser (Marneffe et al. [2006]) to deconstruct sentences' grammatical structure. They argue that grammatical context allows them to more accurately identify the linguistic cues. Consistent with prior research, they find that accruals requiring more estimation are less effective at predicting future earnings.

Predicting word meaning based on its neighboring words (collocation), is generally one of the most common extensions beyond the simple bag-of-words approach. N-gram models estimate this likelihood using Markov chains where memory is limited to a smaller set of words. If this is done by simply counting n-tuples, the process is exponential in the set of words. Much of the work in this area focuses on developing efficient methods for estimating these types of models.[11]

For example, Taddy [2015] has recently proposed an approach based on the general framework of a document language model, where a document is initially transformed into a vector space based on training under a specific objective. In the case relevant here, the objective might be maximizing the likelihood of observing a series of words (within a sentence) around a given word. Taddy proposes that composite likelihoods and Bayes rule can use the local language models to create document classifiers. In this example, the composite likelihood for a sentence is aggregated from the probability of observing each word conditional on the other words. The probabilities of the sentences, under the assumption that they are independent, can be aggregated into the probability of observing the document. Using a training sample to identify the document probabilities for given outcome classes (e.g., sentiment categories), Bayesian inversion can then be used to determine the classification probability for a given document.

5. Measuring Document Similarity

Although much of the emerging literature in textual analysis focuses on sentiment analysis or tone, another useful set of methods in this area provides a means for measuring document similarity. Identifying semantic similarity, which is the basis for document similarity,

[11] An N-gram is a continuous sequence of n words. Also frequently considered are skip-grams where you specify k skips over each set of n words.

is a relatively straightforward task for humans, but is computationally challenging (e.g., why are "cat" and "mouse" related?).

Novel examples of measuring document similarity are provided by Brown and Tucker [2011], who examine changes in MD&A disclosures, Hoberg and Phillips [2015] who focus on 10-K product descriptions to create text-based industry classifications, and Lang and Stice-Lawrence [2015] who compare annual report similarity. These papers use a standard approach taken from the natural language processing and information science literature, a method labeled cosine similarity.

Given two documents d_1 and d_2 that have been collapsed into two vectors x and y of word counts, the cosine similarity measure for i=1 to N words is defined as:

$$\text{cosine similarity}(d_1, d_2) = \frac{\sum_i x_i y_i}{\sqrt{\sum_i x_i^2}\sqrt{\sum_i y_i^2}} \qquad (4)$$

The dot product appearing in the numerator provides a simple measure of similarity, while the denominator scales by the vectors' Euclidean lengths. Geometrically, the measure is the cosine of the angle between the two vectors. Although the numerator provides a measure of similarity, by itself it is difficult to interpret because of the lack of scale. The denominator normalizes the measure, which in this case is useful to make the measure invariant to document length. Although the normalized measure can take on values from -1 to 1, for word counts, which are always non-negative, the measure will range from 0 to 1.

Notice that if we had mean adjusted the counts for each vector in equation (4), the measure would be a simple Pearson correlation.[12] Although the cosine similarity measure is the measure of choice in the natural programming language literature, there seems little reason in business disciplines to use this measure instead of the simple correlation.[13]

Egozi, Markovitch, and Gabrilovich [2011] propose an interesting combination of cosine similarity and latent semantic analysis which uses Wikipedia as a means of establishing context for word classification. Although we are not aware of applications of this technique in accounting and finance, the body of accounting standards and rules in sources such as Generally Accepted Accounting Principles (GAAP), Generally Accepted Auditing Standards (GAAS), and

[12] Brown and Tucker [2011] also note this relation in their footnote 8.

[13] For very large sparse matrices, the cosine similarity can provide some computational advantages. However given current technology, in most cases this advantage is not important. For an interesting discussion of the relation between cosine, correlation, and regression coefficients see Rodgers and Nicewander [1988].

International Financial Reporting Standards (IFRS) would provide a useful context for computationally determining word sense and measuring document similarity.

6. Implementation: Tripwires, Technology, and a Simple Example

In this section, we discuss some of additional areas of caution associated with the implementation of textual analysis and consider various aspects of executing the method. We also provide a simple example of its application.

6.1 WHAT IS A WORD?

Since all textual methods are based on first identifying words, one of the initial steps is to parse each document into a vector of tokens, where tokens are collections of characters occurring between word boundaries.[14] This step can produce a relatively long list which is substantially and meaningfully shortened by selecting only those tokens which map into a list of words. This, in turn, requires the researcher to specify what collections of characters are considered words.

Because word lists have substantial value for hackers attempting to guess passwords, the internet has a wide variety of lists available. If acronyms and proper nouns are included, the lists can exceed one million entries. Loughran and McDonald [2011] develop a word list for business-related textual analysis that is based on the "2of12inf" dictionary.[15] This list includes word inflections but does not include abbreviations, acronyms, or proper nouns. The 2of12inf word list contains more than 80,000 words. For their dictionary, Loughran and McDonald extend this core list by tabulating all tokens in all 10-K and 10-Q variants in the SEC's EDGAR files. All tokens with a frequency count of 100 or more and that are identifiable as words are added to the dictionary. Loughran and McDonald update their dictionary every other year. The only proper noun they include in the list is "Scholes", given the importance and frequency of the term Black-Scholes.

Note that Loughran and McDonald [2011] do not include the single letter words "I" and "a", for two reasons: 1) in most cases they are not important for purposes of analysis, and 2)

[14] Word boundaries are characters such as blank spaces, carriage returns, line feeds, punctuation, or any non-alphanumeric character. The specifics in terms of whether characters such as an underscore are included as part of a word boundary vary across regular expression engines.

[15] The 2of12inf dictionary is documented online at http://wordlist.aspell.net/12dicts-readme-r5/.

such characters are frequently used to enumerate lists.[16] Also, their dictionary includes inflections instead of stemming the words to common lexemes because if the focus is on tone, they have found that using explicit inflections is less error prone than extending a word using stemming (root morpheme + derivational morphemes).

Another aspect of tabulating words is determining what to do with hyphenated tokens. Given that their dictionary provides a count of words appearing in financial disclosures, Loughran and McDonald disambiguate hyphens by looking at the relative likelihood of the two possible outcomes—i.e., the two tokens concatenated without the hyphen, and the two tokens considered separately as words.

6.2 WHAT IS A SENTENCE?

In many studies focusing on readability, the researcher is required to calculate the average number of words per sentence—notably the Fog Index requires this enumeration. That this task is accomplished is usually stated as a simple step in the textual analytic procedures.

For purposes of parsing, the differences between a typical novel and a financial disclosure are extraordinary. In all aspects of textual analysis, this is an important feature to keep in mind. In the case of sentence disambiguation, these differences are crucial. The presence of extensive lists, technical terminology, and other formatting complexities, makes sentence disambiguation especially challenging in accounting disclosures. A typical attempt will first remove abbreviations, headings, and numbers (with decimals), and then assume the remaining periods are sentence terminations. Average words per sentence is then determined by the number of words divided by the number of sentence terminations.[17] Unless some arbitrary level of truncation (like no more than 60 words per sentence) is included, the parsing errors in this case can be extraordinary. Sometimes, researchers might misidentify sentence terminations. Along these lines, Bushee, Gow, and Taylor [2015] highlight that the Perl routine used by Li [2008] to calculate the number of words per sentence is "confounded by punctuation used in

[16] As previously emphasized, all textual analysis tools should be modified in the context of their application. Kim [2013] provides an interesting example where "I" is a useful target in his analysis of self-attribution by CEOs in their media interviews. He finds that CEOs who excessively attribute positive performance to their own skills and negative performance to uncontrollable factors impact the market less and are more likely to be fired.

[17] Alternatively, some will use a complex regular expression that attempts to identify collections of characters exceeding a reasonable length and terminated by punctuation. In our experience, this approach tends to provide relatively unstable measures.

numbers and abbreviations." Thus, they assert that the Fog Index values tabulated by Li [2008] and many other researchers have erroneously low Fog Index values.

The point here is that what seems like a relatively straightforward step in the parsing process can be complex and in many cases very volatile. The computational linguistics literature has a long history of trying to resolve the challenge of disambiguating sentences (see, for example Palmer and Hearst [1994] or Mikheev [2002]). To the extent identifying sentences is a critical aspect of a textual study, the researcher must carefully identify the steps taken to avoid some of the challenges associated with this aspect of parsing. Generic sentence parsing algorithms do not work well on financial documents.

6.3 WHY POSITIVE TONE OR NET TONE IS PROBLEMATIC

Much of the developing literature in textual analysis focuses on the simple distinction between positive and negative information. Given the nuance of language in general and the amorphous nature of the English language, focusing on simple attributes versus subtle affective states is laudable. Besides the obvious dichotomy of positive/negative, some papers also consider the net measure of the two constructs. In the work that we have done in this area, our results have repeatedly reinforced what we observe in reading financial documents. That is, negative words seem unambiguous – rarely does management negate a negative word to make a positive statement.

Positive words, on the other hand, in addition to their positive usage, are just as frequently used to frame a negative statement. Consider a simple case: A careful manager might use 90% positive words in dismissing an employee. As an additional example, the annual report of General Motors (GM), filed on February 28, 2008, highlights that "in 2007, the global automotive industry continued to show strong sales and revenue growth" (p. 48). This is before noting that GM lost $38.7 billion in 2007 (one of the largest annual losses in US corporate history). Although the quick computational response to this issue is to account for negation surrounding positive words, typically negation is far more complex than what will be identified by looking for words like "no" and "not" preceding the target word.

The SEC press release charging BHP Billiton with violating the Foreign Corrupt Practices Act (FCPA) provides another good example of the difficulty in accounting for negation (see http://www.sec.gov/news/pressrelease/2015-93.html). In the press release, the SEC states

"A 'check the box' compliance approach of form over substance is not enough to comply with the FCPA," said Antonia Chion, Associate Director of the SEC's Division of Enforcement. The negation word, "not", appears seven words after the word "compliance". Simply tabulating the count of positive words like "compliance" would misinterpret the sentence's meaning.

The framing of negative information is so frequently padded with positive words that the measured positive sentiment is ambiguous. Although some papers have identified statistically significant effects associated with positive tone (e.g., Jegadeesh and Wu [2013] and Garcia [2013]), Tetlock [2007] and Loughran and McDonald [2011] find little incremental information in positive word lists, which is consistent with the concern about negation of positive words.

6.4 TARGETING SECTIONS IN MANDATED DISCLOSURES

One corpus of textual material that has received considerable attention in the accounting and finance literature on textual analysis is the 10-K annual filings with the SEC. If you are considering a particular textual measure in this context, one of the first suggestions you might receive is to focus your analysis on a specific section of the document. For example, if you are trying to measure uncertainty in the document, some would argue that you should parse out and focus on the MD&A section of the 10-K filing.

At first glance this would appear to be straightforward—i.e., find the segment labeled "Item 7. Management Discussion and Analysis". However, there are obvious qualifiers that must be addressed, e.g., you do not want to identify this phrase in the Table of Contents; you do not want to identify a sentence using this phrase as a reference to the segment; "7" could be a number, a roman numeral, or a word; and the word "Item" may or may not occur. After finding the MD&A heading, then you simply find the subsequent "Item 8" heading and select all content between the two. Parsing tools provide methods to perform these functions fairly accurately. Any error in parsing, however, has the potential to produce extraordinary results. If you misidentify either heading, then the length and content of the MD&A section for this particular document will be substantially misspecified.

How might this happen? Let us list just a few tripwires: (1) prior to about 2002, the 10-K filings are far less structured; (2) many times a segment is mislabeled, e.g. "Item 7" is sometimes listed as "Item 6" (e.g., CIK=1040277, filing date = 20090113); (3) the phrase for MD&A has many variants and is sometimes misspelled (e.g., CIK=1084415, filing date = 20090410); and,

(4) the MD&A section might be reported as exhibit 13 and will not appear in the main body of the filing.

An additional problem for researchers focusing on only one section is that companies can shift content between sections. If the discussion in the MD&A section is less extensive, the footnotes for various accounting items could be more extensive and provide additional information. Important compensation information might be in the 10-K or it could be in a proxy statement. In sum, what seems like an obvious segmentation of the document, computationally is not.

6.5 LEVELS VERSUS DIFFERENCES

To an extent, where there is a time series of firm specific documents, some of the issues caused by misclassification from using a poorly designed off-the-shelf dictionary can be mitigated in word-count methods by differencing. For example, the effect of frequently occurring misclassified words like *crude, vice,* and *mine* will be reduced by differencing. This helps to explain why apparently inferior dictionaries perform as well as the LM word lists in many settings when unexpected tone (i.e., the difference between the two tone measures) is the key variable.

It is important to note that when document tone is differenced, the importance of Zipf's law is mitigated. Zipf's law applies to levels, but the distribution of changes in tone can be driven by numerous words that are not the most frequently occurring overall words.

Should researchers examine tone levels or tone differences? If we are considering earnings conference calls or 10-K filings, are we focused on the variation in the cross-section or the changes within a single firm across time? In the case of a 10-K, differencing would imply a scenario where the reader was making a year-to-year comparison of tone versus comparing with the cross-section. The majority of textual based studies focus on levels, but some (see, for example, Feldman et al. [2010]) have used differences. Brown and Tucker [2011] provide a clear example where they analyze differences in the MD&A for a given firm. They find firms with more changes in their MD&A section also have larger economic changes, have a higher magnitude of stock price response to the 10-K filing, and that this response has declined in the past decade. Clearly, the economic logic underlying the use of the documents should dictate the structure of the experimental design.

6.6 HOW TO IMPLEMENT

6.6.1. Programming Languages for Textual Analysis. Programming languages are a matter of religion, but more experienced coders are less likely to proclaim a dominant platform. Ignoring the very early history in the area, Perl was traditionally considered the programming language of choice when analyzing text. Perl's popularity in the past few years has languished, while Python has become a generic solution to common programming tasks. Both of these software platforms have widely available pre-packaged solutions for various parsing tasks such as HTML removal, sentence parsing, or word parsing.[18] All of the major statistical software packages, Stata, SAS, SPSS, and R, also are available as platforms that are very capable of analyzing text.

The key component for any programming language to be used in parsing text is the availability of a regular expression processor within the language. Regular expressions, or "regex", provides for efficient pattern searches within text. For example, one might use the expression "(?s:<Table(.*?)>(.*?)</Table>)" to identify all of the material contained in a table within an SEC document, or the sequence "\b[-+\(]?\p{Sc}?[-+(]?[\d,.]+" might be used to identify collections of numbers. Specifying a single regex to identify complex sequences rarely produces a solution that will always correctly parse the document, and much time must be spent tuning the algorithm to the specific corpus of text being considered.

The danger of using programs with prepackaged parsing tools is not unlike the concerns in using such programs for statistics. That is, statistical packages allow users to apply complex statistical solutions in cases where they might not fully appreciate the requisite assumptions for their application. Similarly, if we have a package that parses a document for sentences, we might not fully appreciate the challenges of accurately disambiguating capitalized words and abbreviations, especially in the context of financial documents which tend to have a relatively complex formatting structure.

One simple example of this is the Fathom package in Perl which provides a means of generating syllable counts for words. Using a list of more than 40,000 words with pre-identified syllable counts, we found that the Fathom package was accurate in only about 75% of the cases. Using the Talburt method, another popular approach, had about the same level of accuracy. By

[18] For example, Lingua in Perl and the Natural Language Toolkit (NLTK) in Python.

tuning the code to the text being tested, we were able to easily get the accuracy to over 90%. Gains beyond that are essentially writing code for individual exceptions.

In this paper, we underscore the problems with using technologies not developed within the context of the media being studied. Similarly using generic prepackaged programs to parse business-related documents, which in some cases can contain relatively complex formatting structures, creates significant uncertainty about the accuracy of the textual measures. However, a norm where each researcher produces independent code does not seem efficient and is equally prone to error, or at least difficulties in replication. We suggest using a solution that has become relatively common in software development. Much as our profession has shared many core datasets, we are establishing a repository with common routines used in textual analysis (Notre Dame Software Repository for Accounting and Finance at http://sraf.nd.edu). Although we will initially develop the code in Python, the site will be open to alternative solutions. If successful, this repository will provide a systematic collection of open-source software that will standardize specific applications in accounting and finance.

6.6.2. A Simple Example. As discussed before, one of the most straightforward applications of textual analysis is when the researcher can identify an unambiguous word or phrase and simply tabulate the presence of this phrase in a financial document. We provide an example of this simple case. For our example, we hypothesize that firms using the phrase "non-GAAP" in 10-K filings (including all 10-K variants) are more likely to have higher subsequent stock return volatility. As in Loughran and McDonald [2014], we use the market model root-mean-square for the post-filing trading days [6, 28] as a measure of uncertainty in the information environment.

In this case the "textual analysis" is straightforward—we have a target phrase which is well defined, is not likely to be misidentified (e.g., it does not frequently show up in the name of a company or as a homonym), and is very easy to parse out of the document. Because the researcher is most likely going to parse the documents repeatedly in developing their software, a first step would be to download all of the SEC 10-K documents to local storage. The SEC Master File for filings provides, for each quarter and year, a master list of filings with their server file path, which allows this process to be computationally implemented. Among the various filings associated with the 10-K, the ".txt" version of the file contains all of the

information associated with the filing. To facilitate using these files, the researcher will typically want to exclude all extraneous attributes such as HTML, XBRL, and embedded binaries (e.g., graphics, pdf files, and Microsoft Excel files).

In any application, a decision must be made as to whether tables or exhibits are included in the analysis. Identifying tables can be challenging because many firms (especially prior to 2005) embed all paragraphs, including pure text, in table markings (e.g., <Table>). We consider each segment of a document demarcated by HTML as a table by counting the number of alphabetic versus numeric characters. We label HTML-identified tables as a "true" table for segments with more than 10% numeric characters.

Once the document has been cleansed of these components, in this example we simply look for the occurrence of the phrase "non-GAAP". In this particular case we are not concerned about whether the token is upper or lower case, so the search ignores this attribute. In a more complex application where words are being classified, we would first use a regular expression to create a list of tokens, i.e., character collections preceded and followed by word boundaries (e.g., blanks, line feeds, punctuation, or carriage returns). We would then use the dictionary we discussed previously to tabulate word frequencies.

Our assumption for this empirical example is that managers direct investor's attention to non-GAAP numbers when non-GAAP results paint a rosier picture of the firm's prospects. For example, if the firm has a heavy debt load, managers may elect to highlight positive trends in non-GAAP EBITDA, thereby glossing over the company's inability to generate cash flows much above their interest expense. This discussion away from GAAP numbers should make valuations by investors more problematic; thus increasing the firm's stock return volatility.

Using 10-K filings from 1994-2011, we find that the term non-GAAP occurs in only about 7.7% of the filings. The regression results, where post-filing date excess return volatility is the dependent variable, are reported in Table 1. The first column replicates the regression results of Loughran and McDonald [2014] using only their control variables. In the second column, we include a dummy variable for cases where non-GAAP appears at least once. The coefficient on the *non-GAAP* dummy variable is 0.054 with a *t*-statistic of 2.64. The results from this simple exercise suggest that the use of the term *non-GAAP* is associated with increased post-filing stock

return volatility. A recent *Wall Street Journal* article argues that the use of non-GAAP accounting in IPO offerings could confuse investors.[19]

It is important to note that even if a researcher correctly tabulates the frequency of "non-GAAP" in an annual report, then we still have the problem that management's choice to report non-GAAP numbers is completely voluntary. Thus, it is not clear if we are measuring the effect of non-GAAP usage or the shocks to the firm's business that motivates it to highlight non-GAAP reporting in the first place.

7. Areas for Future Research in Textual Analysis

Disentangling the role of firm-level complexity from readability is problematic. A complex firm (i.e., multiple diverse divisions, opaque corporate structure, and/or hard to understand business models) might be expected to produce business documents that are more difficult to read solely due to the nature of their business operations. Much of the prior literature has used rather crude measures to control for firm-specific complexity.

For example, You and Zhang [2009] use the median 10-K word count to classify firms into low complexity and high complexity groups. Other researchers have used the number of business segments or a Herfindahl Index based on segment revenue to capture firm-level complexity (see Huang et al. [2014] and Loughran and McDonald [2014]). Counter examples, however, are easy to generate. Large multinational firms like Coca-Cola will consistently have above median 10-K word counts due to the need to describe their various markets, yet will have a low complexity business model. Some firms could operate in only one business segment, yet would be classified as complex by readers of their business communications. Clearly business complexity comes in many forms.

In addition to the challenges of separating out the concepts of business complexity and readability, the meaning of readability in the context of business documents is not clear. We would suggest that future research should focus on the broader concept of information complexity and not consider the concept of readability, which is confounded by its historical usage. Although measures such as the log of file size provide reasonable proxies for this broader concept of information complexity in 10-K filings, this measure might not be useful in shorter, less varied business documents such as press releases, earnings conference calls, or 8-K

[19] See http://www.wsj.com/articles/tailored-accounting-at-ipos-raises-flags-1420677431.

announcements. We need to develop measures of information complexity that are not simply a function of document size. Studies that choose to focus on readability in spite of the concerns we have identified, must carefully define what is meant by the concept and how a specific readability measure balances technical precision in writing with simplicity.

As we discussed in an earlier section, term weighting has the potential to increase the power of textual methods, but lacking theoretical motivation or independent verification, provides the researcher with potentially too many degrees of freedom in selecting an empirical specification. Hopefully future research will provide a structured analysis that provides an objective basis for specifying a particular weighting scheme in textual applications.

As noted earlier, much of the literature in accounting and finance uses a bag of words approach to measure document sentiment. In the context of business applications, is there information that might be relevant if we adopt the methods of deep parsing—where sentences are broken down into meaningful components—to assess a document? Can the methods of "Cloud Robotics" and "Deep Learning" (see, for example, Pratt [2015]), where machine learning is augmented by enormous cloud-based training sets, be adapted to capture deeper meaning and context in business text? Many of these more complex methods potentially add more noise than signal. Researchers introducing new techniques to the literature must bear the burden of carefully explaining the method, considering its power in their specific application, and providing transparent results.

Although a given study will raise suspicions if it arbitrarily defines its own sentiment words, there are clearly potential words that might impact sentiment measures within the context of different bodies of text. For example, Larcker and Zakolyukina [2012] create their own word lists to examine conference calls including negative words not in the LM list such as *atrocious*, *barbarous*, *farcical*, *idiot*, and *wonky*. Can the LM word lists, which were developed in the context of 10-K filings, be adapted to alternative modes of business writing such as newspaper articles or conference call transcripts? In many instances, the approach used by Allee and DeAngelis [2015], where they transparently modify the LM word lists to fit their specific application, would seem useful.

Many textual analysis studies have focused on the simple positive/negative dichotomy of sentiment analysis. We argue that tests for positive sentiment have low power. LM also created word lists for "uncertainty", "litigious", "strong modal", and "weak modal" words. These lists

could provide an additional means of parsing sentiment and there are other systematic word groupings that might produce useful targets.

An economic hypothesis that could be examined is whether or not managers using high levels of uncertain or weak modal (e.g., weasel) words during conference calls experience worse subsequent stock or operating performance. Additionally, researchers could examine the sentiment of the initial media coverage on an announced acquisition. Articles with higher uncertainty sentiment could predict significantly lower probabilities of completing the merger. This would provide another link between sentiment in media articles and subsequent economic outcomes.

Gentzkow and Shaprio [2010] examine media slant in newspapers by using the Congressional Record to creatively produce a list of phrases commonly used in the speeches of Democrat versus Republican members of Congress. The authors then categorize over 400 daily newspapers by political slant using the most partisan phrases in the language of elected officials. Thus, newspapers that more frequently use phrases like *trade deficit, oil companies, Artic refuge,* and *minimum wage* would be classified as more liberal. Conversely, papers using the phrases *war on terror, private property, human embryos,* and *retirement accounts* would have a more Republican slant. Perhaps, early on, we can identify managers of firms destined for success or failure based on their use of language in communications. Do great leaders share similar communication attributes?

Although parsing SEC filings for sentiment is far more challenging than parsing the typical novel, parsing social media is even more challenging. The use of slang, emoji, and sarcasm, and the constantly changing vocabulary on social media makes the accurate classification of tone difficult. And yet social media is a central source of emerging information, with some channels focused on business activities (e.g., http://stocktwits.com). Hopefully methods can be developed that are better able to capture the information in this very noisy yet rich source of data.

We have only focused on textual analysis in the English language. Other languages present their own advantages and challenges. Tsarfaty et al. [2013] and Das and Banerjee [2012] provide examples of challenges from alternative languages. For example, the German language is much more structured than English, but also suffers from syncretism, which is the case where a word form serves multiple grammatical purposes (e.g., in English, “bid” is both present and

past tense). Non-configurational languages, such as Hungarian, where word order is less important, makes syntactical analysis difficult. Annotations for French are frequently inadequate due to the language's diversity of word forms. The initial step of chunking a document into words is more challenging in Chinese and Japanese where text is usually not demarcated by inter-word spaces. Presumably, linguistic typologists can provide useful insights into which methods are most appropriate for which languages.

8. Conclusion

Information plays a central role in how accountants document a firm's operations and how financial markets assess value. A generation of researchers has carefully considered how the quantitative data in accounting and finance refine our understanding of business and financial markets. Almost all quantitative data in this arena is contextualized by textual information which we are just now beginning to explore for deeper insights. In this paper we have tried to review the growing literature on textual analysis in accounting and finance, discuss the most commonly applied methods, and report some of the tripwires associated with the methods.

We can summarize what we have learned from our prior work and this survey of the related literature in five points:

(1) The traditional concept of readability—where word and sentence length are important determinants of grade-level comprehension—does not map well into determining the effectiveness of business documents as information conduits. The most common measure of readability, the Fog Index, is ineffective because: (a) measuring sentence length is error prone in business documents, (b) most multisyllabic words in business documents are easily understood, and most importantly, (c) the logic that a financial document targeted at a lower grade level is more effective is questionable. Loughran and McDonald [2014], in addition to documenting the weaknesses of the Fog Index, show that financial jargon is positively associated with their measures of information integration. This result contradicts the traditional interpretation of readability and suggests that words targeted at a financially sophisticated audience can make documents more effective. We believe the essential concept of readability that we have borrowed from other disciplines would be better captured if instead it was examined under the

framework of information complexity. Information complexity would encompass both document complexity and business complexity.

(2) Zipf's law documents the fact that in any non-pathological list of words, a very small number of words will dominate the frequency counts. This property of word distributions creates a research environment where seemingly innocent word misclassifications do not simply add small amounts of random noise to your results and can produce outliers that drive spurious results.

(3) From our experience, the best way to avoid the numerous and substantial tripwires in textual analysis is to carefully consider the ability of a program, word list, and statistical method to work effectively in the specific context of application. Avoid using word lists and algorithms derived in the context of other disciplines unless they are first proven to be effective within the domain of your research. Avoid black boxes where you feed a document into a generic program and it produces your results.

(4) In publicizing your research, error on the side of transparency. Parsing methods must be documented in detail, especially for cases where the underlying documents have complex structures. The dictionary used to determine when a token is a word must be identified. Classification methods, whether simple bag-of-words approaches or more complex outcomes from machine learning tools, must reveal the words and topics driving the results. Unless a study can convincingly resolve the problems of negation, positive sentiment is best left untested. The fact that many journals now offer web appendices for background detail makes such transparency even more practical.

(5) Provide a description of your method that makes your research replicable. The literature would also benefit from the posting of programs that could run texts of other researchers through the same engine. We are initiating a software repository that attempts to begin addressing this issue (Notre Dame Software Repository for Accounting and Finance at http://sraf.nd.edu).

In sum, all textual studies in accounting and finance must consider carefully the transparency and replicability of their results. The more complex the method, the more transparency must be emphasized. In hindsight, as the profession has begun to explore the relatively new method of textual analysis, many methodological choices would likely be different when considered in the

context of what we now know. Some central empirical results concerning sentiment, earnings, management choices, and stock prices should be re-evaluated. With increasing computational power and an explosion of digital text available for research there is much yet to be done.

References

AHERN, K., AND D. SOSYURA. "Who Writes the News? Corporate Press Releases during Merger Negotiations." *Journal of Finance* 69 (2014): 241-291.

ALLEE, K., AND M. DEANGELIS. "The Structure of Voluntary Disclosure Narratives: Evidence from Tone Dispersion." *Journal of Accounting Research* 53 (2015): 241-274.

ANTWEILER, W., AND M. FRANK. "Is All that Talk just Noise? The Information Content of Internet Stock Message Boards." *Journal of Finance* 59 (2004): 1259-1294.

BERELSON, B. *Content Analysis in Communication Research*, Glencoe, IL: The Free Press (1952).

BIDDLE, G.; G. HILARY; AND R. VERDI. "How does Financial Reporting Quality relate to Investment Efficiency?" *Journal of Accounting and Economics* 48 (2009): 112–131.

BLEI, D.; A. NG; AND M. JORDAN. "Latent Dirichlet Allocation." *Journal of Machine Learning Research* 3 (2003): 993-1022.

BLOOMFIELD, R. "Discussion of Annual Report Readability, Current Earnings, and Earnings Persistence." *Journal of Accounting and Economics* 45 (2008): 248–252.

BODNARUK, A.; T, LOUGHRAN; AND B. MCDONALD. "Using 10-K Text to Gauge Financial Constraints." *Journal of Financial and Quantitative Analysis* 50 (2015): 623-646.

BONSALL, S. B.; A. J. Leone; AND B. P. MILLER. "A Plain English Measure of Financial Reporting Readability." Working paper, Ohio State University, 2015.

BONSALL, S. B. AND B. P. MILLER. "The Impact of Narrative Disclosure Readability on Bond Ratings and Rating Agency Disagreement." Working paper, Ohio State University, 2014.

BOUKUS, E., AND J. ROSENBERG. "The Information Content of FOMC Minutes." Working paper, Federal Reserve Bank of New York, 2006.

BRATTEN, B.; C. A. GLEASON; S. LAROCQUE; AND L. F. MILLS. "Forecasting Tax Expense: New Evidence from Analysts." Working paper, University of Notre Dame, 2014.

BROWN, S., AND J. W. TUCKER. "Large-Sample Evidence on Firms' Year-over-Year MD&A Modifications." *Journal of Accounting Research* 49 (2011): 309-346.

BUEHLMAIER, M., AND T. WHITED. "Looking for Risk in Words: A Narrative Approach to Measuring the Pricing Implications of Finance Constraints." Working paper, University of Rochester, 2014.

BUEHLMAIER, M., AND J. ZECHNER. "Slow-moving Real Information in Merger Arbitrage." Working paper, University of Hong Kong, 2013.

BURKE, K. "The Rhetoric of Hitler's 'Battle'." *The Southern Review* 5 (1939): 1-21.

BUSHEE, B. J.; I. D. GOW; AND D. J. TAYLOR. "Linguistic Complexity in Firm Disclosures: Obfuscation or Information?" Working paper, University of Pennsylvania, 2015.

The Catholic Encyclopedia, Volume 4, 1908.

CHEN, H.; P. DE; Y. HU; AND B. H. HWANG. "Wisdom of Crowds: The Value of Stock Opinions Transmitted through Social Media." *Review of Financial Studies* 27 (2014): 1367-1403.

CHEN, J. V., AND F. LI. "Estimating the Amount of Estimation in Accruals." Working paper, University of Michigan, 2013.

COVAL, J., AND T. SHUMWAY. "Is Sound just Noise?" *Journal of Finance* 61 (2001): 1887-1910.

CROSSNO, P.; A. WILSON; T. SHEAD; AND D. DUNLAVY. "Topicview: Visually Comparing Topic Models of Text collections." In *Tools with Artificial Intelligence (ICTAI), 1022 23rd IEEE International Conference* (2011): 936-943.

DAS, S. R. "Text and Context: Language Analytics in Finance." *Foundations and Trends in Finance* 8 (2014): 145-261.

DAS, S., AND S. BANERJEE. "Pattern Recognition Approaches to Japanese Character Recognition." *Advances in Computer Science, Engineering and Applications* 166 (2012): 83-92.

DAS, S. R., AND M. Y. CHEN. "Yahoo! for Amazon: Sentiment Extraction from Small Talk on the Web." *Management Science* 53 (2007): 1375-1388.

DAVIS, A. K.; W. GE; D. MATSUMOTO; AND J. L. ZHANG. "The Effect of Manager-specific Optimism on the Tone of Earnings Conference Calls." *Review of Accounting Studies* 20 (2015): 639-673.

DAVIS, A. K.; J. M. PIGER; AND L. M. SEDOR. "Beyond the Numbers: Measuring the Information Content of Earnings Press Release Language." *Contemporary Accounting Research* 29 (2012): 845-868.

DAVIS, A. K., AND I. TAMA-SWEET. "Managers' use of Language across Alternative Disclosure Outlets: Earnings Press Releases versus MD&A." *Contemporary Accounting Research* 29 (2012): 804-837.

DE FRANCO, G.; O. HOPE; D. VYAS; AND Y. ZHOU. "Analyst Report Readability." *Contemporary Accounting Research* 32 (2015): 76-104.

DORAN, J. S.; D. R. PETERSON; AND S. M. PRICE. "Earnings Conference Call Content and Stock Price: The Case of REITs." *The Journal of Real Estate Finance and Economics* 45 (2012): 402-434.

DOUGAL, C.; J. ENGELBERG; D. GARCIA; AND C. A. PARSONS. "Journalists and the Stock Market." *Review of Financial Studies* 25 (2012): 639-679.

EGOZI, O.; S. MARKOVITCH; AND E. GABRILOVICH. "Concept-Based Information Retrieval Using Explicit Semantic Analysis. *ACM Transactions of Information Systems* 29 (2011): 8-32.

ERTUGRUL, M.; J. LEI; J. QIU; AND C. WAN. "Annual Report Readability, Tone Ambiguity, and the Cost of Borrowing." *Journal of Financial and Quantitative Analysis* (2015): forthcoming.

FRAZIER, K. B.; R. W. INGRAM; AND B. M. TENNYSON. "A Methodology for the Analysis of Narrative Accounting Disclosures." *Journal of Accounting Research* 22 (1984): 318-331.

FELDMAN, R.; S. GOVINDARAJ; J. LIVNAT; AND B. SEGAL. "Management's Tone Change, Post Earnings Announcement Drift and Accruals." *Review of Accounting Studies* 15 (2010): 915-953.

FERRIS, S. P.; G. HAO; AND M. LIAO. "The Effect of Issuer Conservatism on IPO Pricing and Performance." *Review of Finance* 17 (2013): 993-1027.

GARCIA, D. "Sentiment during Recessions." *Journal of Finance* 68 (2013): 1267-1300.

GENTZKOW, M., AND J. M. SHAPIRO. "What Drives Media Slant? Evidence from U.S. Daily Newspapers." *Econometrica* 78 (2010): 35-71.

GUAY, W.; D. SAMUELS; AND D. TAYLOR. "Guiding Through the Fog: Financial Statement Complexity and Voluntary Disclosure." Working paper, University of Pennsylvania, 2015.

GURUN, U. G., AND A. W. BUTLER. "Don't believe the Hype: Local Media Slant, Local Advertising, and Firm Value." *Journal of Finance* 67 (2012): 561-598.

HANLEY, K. W., AND G. HOBERG. "The Information Content of IPO Prospectuses." *Review of Financial Studies* 23 (2010): 2821-2864.

HENRY, E. "Are Investors Influenced by how Earnings Press Releases are Written?" *Journal of Business Communication* 45 (2008): 363-407.

HESTON, S. L., AND N. SINHA. "News versus Sentiment: Predicting Stock Returns from News Stories." Working paper, University of Maryland, 2015.

HILLERT, A.; A. NIESSEN-RUENZI; AND S. RUENZI. "Mutual Fund Shareholder Letter Tone–Do Investors Listen?" Working paper, University of Mannheim, 2014.

HOBERG, G., AND G. PHILLIPS. "Text-based Network Industries and Endogenous Product Differentiation." *Journal of Political Economy* (2015): forthcoming.

HOFMANN, T. "Unsupervised Learning by Probabilistic Latent Semantic Analysis." *Machine Learning* 42 (2001): 177-196.

HUANG, A.; R. Lehavy; A. ZANG; AND R. ZHENG. "Analyst Information Discovery and Interpretation Roles: A Topic Modeling Approach." Working paper, University of Michigan, 2015.

HUANG, A.; A. ZANG; AND R. ZHENG. "Evidence on the Information Content of Text in Analyst Reports." *The Accounting Review* 89 (2014): 2151-2180.

HUANG, X.; S. H. TEOH; AND Y. ZHANG. "Tone Management." *The Accounting Review* 89 (2014): 1083-1113.

JEGADEESH, N., AND D. WU. "Word Power: A New Approach for Content Analysis." *Journal of Financial Economics* 110 (2013): 712-729.

JONES, M. J., AND P. A. SHOEMAKER. "Accounting Narratives: A Review of Empirical Studies of Content and Readability." *Journal of Accounting Literature* 13 (1994): 142–184.

KEARNEY, C., AND S. LIU. "Textual Sentiment in Finance: A Survey of Methods and Models." *International Review of Financial Analysis* 33 (2014): 171-185.

KIM, Y. H. "Self Attribution Bias of the CEO: Evidence from CEO Interviews on CNBC." *Journal of Banking & Finance* 27 (2013): 2472-2489.

KLARE, G. "The Measurement of Readability." Ames, IA: Iowa University Press (1963).

KOTHARI, S. P.; X. LI; AND J. E. SHORT. "The Effect of Disclosures by Management, Analysts, and Business Press on Cost of Capital, Return Volatility, and Analyst Forecasts: A Study using Content Analysis." *The Accounting Review* 84 (2009): 1639-1670.

LANG, M., AND L. STICE-LAWRENCE. "Textual Analysis and International Financial Reporting: Large Sample Evidence." *Journal of Accounting and Economics* 60 (2015): 110-135.

LARCKER, D. F., AND A. A. ZAKOLYUKINA. "Detecting Deceptive Discussions in Conference Calls." *Journal of Accounting Research* 50 (2012): 495-540.

LAWRENCE, A. "Individual Investors and Financial Disclosure." *Journal of Accounting & Economics* 56 (2013): 130–147.

LEHAVY, R.; F. LI; AND K. MERKLEY. "The Effect of Annual Report Readability on Analyst following and the Properties of their Earnings Forecasts." *The Accounting Review* 86 (2011): 1087–1115.

LEUZ, C., AND C. SCHRAND. "Disclosure and the Cost of Capital: Evidence from Firms' Responses to the Enron Shock." Working paper, University of Chicago, 2009.

LEUZ, C., AND P. WYSOCKI. "The Economics of Disclosure and Financial Reporting Regulation: Evidence and Suggestions for Future Research." *Journal of Accounting Research* (2016): forthcoming.

LEWIS, N. R.; L. D. PARKER; G. D. POUND; AND P. SUTCLIFFE. "Accounting Report Readability: The Use of Readability Techniques. *Accounting and Business Research* 16 (1986): 199-213.

LI, F. "Annual Report Readability, Current Earnings, and Earnings Persistence." *Journal of Accounting and Economics* 45 (2008): 221–247.

LI, F. "Textual Analysis of Corporate Disclosures: A Survey of the Literature." *Journal of Accounting Literature* 29 (2010a): 143-165.

LI, F. "The Information Content of Forward-looking Statements in Corporate Filings—A Naïve Bayesian Machine Learning Approach." *Journal of Accounting Research* 48 (2010b): 1049-1102.

LI, J., AND X. ZHAO. "Complexity and Information Content of Financial Disclosures: Evidence from Evolution of Uncertainty Following 10-K Filings." Working paper, University of Texas (Dallas), 2014.

LIU, B., AND J. J. MCCONNELL. "The Role of the Media in Corporate Governance: Do the Media Influence Managers' Capital Allocation Decisions?" *Journal of Financial Economics* 110 (2013): 1-17.

LOUGHRAN, T., AND B. MCDONALD. "When is a Liability not a Liability? Textual analysis, Dictionaries, and 10-Ks." *Journal of Finance* 66 (2011): 35-65.

LOUGHRAN, T., AND B. MCDONALD. "IPO First-day Returns, Offer Price Revisions, Volatility, and Form S-1 Language." *Journal of Financial Economics* 109 (2013): 307-326.

LOUGHRAN, T., AND B. MCDONALD. "Measuring Readability in Financial Disclosures." *Journal of Finance* 69 (2014): 1643-1671.

LOUGHRAN, T., AND B. MCDONALD. "The Use of Word Lists in Textual Analysis." *Journal of Behavioral Finance* 16 (2015): 1-11.

LOUGHRAN, T.; B. MCDONALD; AND H. Yun. "A Wolf in Sheep's Clothing: The Use of Ethics-Related Terms in 10-K Reports." *Journal of Business Ethics* 89 (2009): 39-49.

LUNDHOLM, R. J.; R. ROGO; AND J. ZHANG. "Restoring the Tower of Babel: How Foreign Firms Communicate with US Investors." *The Accounting Review* 89 (2014): 1453-1485.

MANNING, C. D., AND H. SCHÜTZE. Foundations of Statistical Natural Language Processing. Cambridge, MA: MIT Press (2003).

MARCUS, M.; B. SANTORINI; AND M. A. MARCINKIEWICZ. "Building a Large Annotated Corpus of English: The Penn Treebank." *Computational Linguistics* 19 (1993): 313-330.

MARNEFFE, M.; B. MACCARTNEY; AND C. MANNING. "Generating Typed Dependency Parses from Phrase Structure Parses. *Proceedings of LREC* 6 (2006): 449-454.

MATSUMOTO, D.; M. PRONK; AND E. ROELOFSEN. "What makes Conference Calls Useful? The Information Content of Managers' Presentations and Analysts' Discussion Sessions." *The Accounting Review* 86 (2011): 1383-1414.

MAYEW, W. J., AND M. VENKATACHALAM. "The Power of Voice: Managerial Affective States and Future Firm Performance." *Journal of Finance* 67 (2012): 1-43.

MCLAUGHLIN, G. "SMOG Grading: A New Readability Formula." *Journal of Reading* 12 (1969): 639-646.

MIKHEEV, A. "Periods, Capitalized Words, etc." *Computational Linguistics* 28 (2002): 289–316.

MILLER, B. P. "The Effects of Reporting Complexity on Small and Large Investor Trading." *The Accounting Review* 85 (2010): 2107-2143.

MOSTELLER, F., AND D. WALLACE. Inference and Disputed Authorship: The Federalist. Addison-Wesley, Reading, MA (1964).

PALMER, D., AND M. A. HEARST. "Adaptive Sentence Boundary Disambiguation." *Proceedings of the Fourth ACL Conference on Applied Natural Language Processing*, (1994): 78–83.

PRATT, G. "Is a Cambrian Explosion Coming for Robotics?" *Journal of Economic Perspectives*, 29 (2015): 51-60.

PRICE, S. M.; J. S. DORAN; D. R. PETERSON; AND B. A. BLISS. "Earnings Conference Calls and Stock Returns: The Incremental Informativeness of Textual Tone." *Journal of Banking & Finance* 36 (2012): 992-1011.

PURDA, L., AND D. SKILLICORN. "Accounting Variables, Deception, and a Bag of Words: Assessing the Tools of Fraud Detection." *Contemporary Accounting Research* (2015): forthcoming.

RENNEKAMP, K. "Processing Fluency and Investors' Reactions to Disclosure Readability." *Journal of Accounting Research* 50 (2012): 1319–1354.

ROGERS, J. L., AND W. NICEWANDER. "Thirteen Ways to Look at the Correlation Coefficient." *The American Statistician* 42 (1988): 59-66.

ROGERS, J. L.; A. VAN BUSKIRK; AND S. ZECHMAN. "Disclosure Tone and Shareholder Litigation." *The Accounting Review* 86 (2011): 2155-2183.

SALTON, G. AND C. BUCKLEY. "Term-Weighting Approaches in Automatic Text Retrieval." *Information Processing & Management* 24 (1988): 513-523.

SOLOMON, D. H. "Selective Publicity and Stock Prices." *Journal of Finance* 67 (2012): 599-638.

SOLOMON, D. H.; E. SOLTES; AND D. SOSYURA. "Winners in the Spotlight: Media Coverage of Fund Holdings as a Driver of Flows." *Journal of Financial Economics* 113 (2014): 53-72.

TADDY, M. "Document Classification by Inversion of Distributed Language Representations." *Proceedings of the 53rd Meeting of the Association for Computations Linguistics* (2015).

TENNYSON, B. M.; R. W. INGRAM; AND M. T. DUGAN. "Assessing the Information Content of Narrative Disclosures in Explaining Bankruptcy." *Journal of Business Finance & Accounting* 17 (1990): 391-410.

TETLOCK, P. C. "Giving Content to Investor Sentiment: The Role of Media in the Stock Market." *Journal of Finance* 62 (2007): 1139-1168.

TETLOCK, P. C.; M. SAAR-TSECHANSKY; AND S. MACSKASSY. "More than Words: Quantifying Language to Measure Firms' Fundamentals." *Journal of Finance* 63 (2008): 1437-1467.

TSARFATY, R.; D. SEDDAH; S. KÜBLER; AND J. NIVRE. "Parsing Morphologically Rich Languages: Introduction to the Special Issue." *Association for Computational Linguistics* 39 (2013): 15-22.

TWEDT, B., AND L. REES. "Reading between the Lines: An Empirical Examination of Qualitative Attributes of Financial Analysts' Reports." *Journal of Accounting and Public Policy* 31 (2012): 1-21.

WILLIAMS, C. B. "Mendenhall's Studies of Word-length Distribution in the Works of Shakespeare and Bacon." *Biometrika* 62 (1975): 207-212.

YOU, H., AND X. ZHANG. "Financial Reporting Complexity and Investor Underreaction to 10-K Information." *Review of Accounting Studies* 14 (2009): 559–586.

ZOBEL, J. AND A. MOFFAT. "Exploring the Similarity Space." *ACM SIGIR Forum* 32 (1998): 18-34.

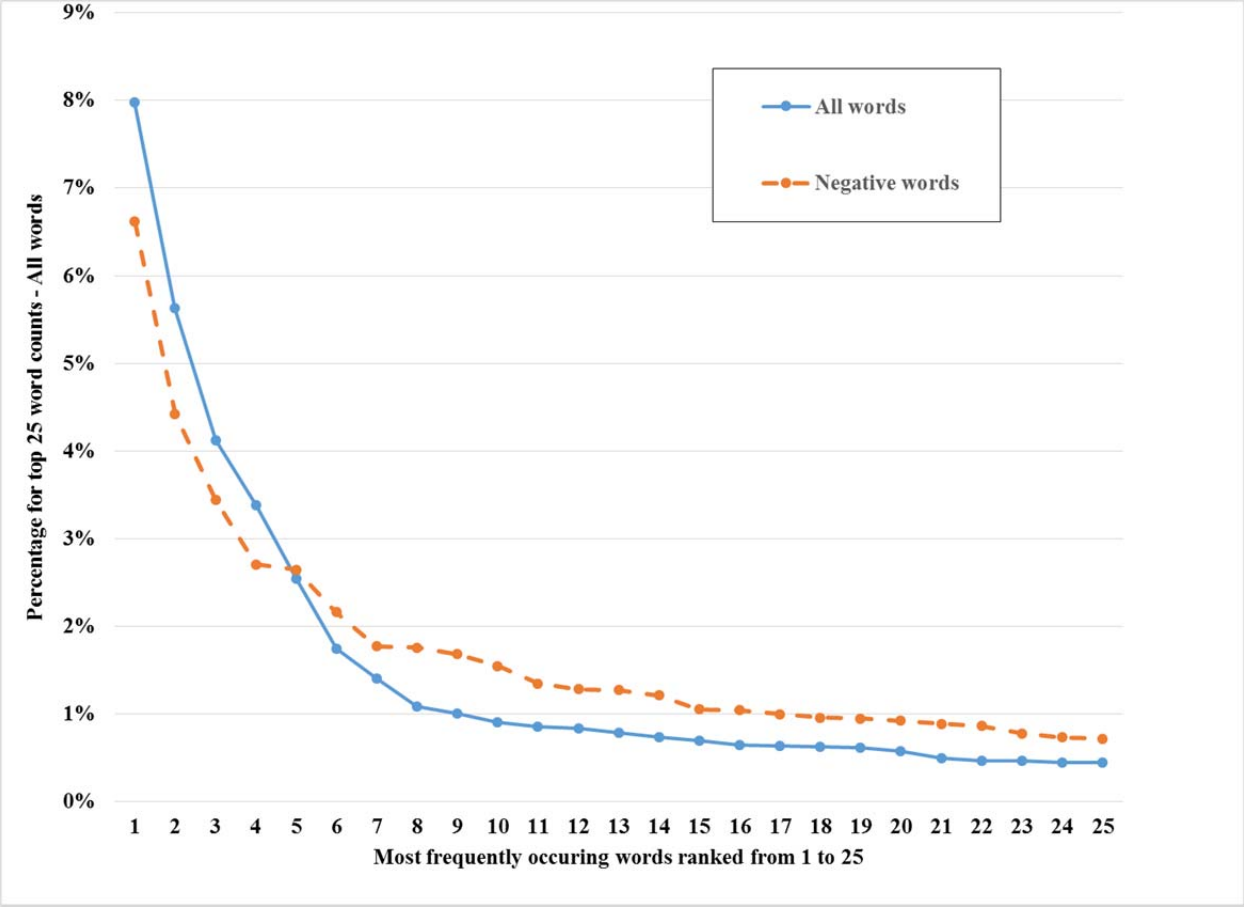

Fig. 1.—The figure shows the proportions for the top 25 most frequently occurring words in all 10-K/Q type SEC filings over the period 1994-2012 for both "All words" and for "Negative words" (using Loughran-McDonald word lists). The denominator for "All words" is the sum of all word counts across all 10-K/Q filings, while for "Negative words" it is the sum of all negative word counts.

TABLE 1

Use of "non-GAAP" in 10-K Filings and Market Model Post Filing Date Root Mean Square Error

Dep. Var. = *PostFilingDateRMSE*	(i)	(ii)
Non-GAAP dummy		0.054***
		(2.64)
Log(File size)	0.073***	0.072***
	(4.60)	(4.57)
Pre-filing market model alpha	-0.898***	-0.898***
	(-4.06)	(-4.06)
Pre-filing market model RMSE	0.536***	0.535***
	(11.89)	(11.90)
Abs(filing period return)	5.046***	5.048***
	(17.56)	(17.59)
Log(size in $ millions)	-0.117***	-0.118***
	(-5.91)	(-5.87)
Log(book-to-market)	-0.140***	-0.140***
	(-2.52)	(-2.53)
NASDAQ dummy	0.264***	0.265***
	(3.38)	(3.38)
Intercept	1.537***	1.491***
	(6.29)	(5.96)
Industry dummies	Included	Included
Year dummies	Included	Included
No. of observations	66,707	66,707
Adj. R^2	46.96%	46.97%

This table presents regressions that test the impact of using the term "non-GAAP" in 10-K filings. The sample and control variables are defined in Loughran and McDonald [2014]. Reported *t*-statistics are adjusted for clustering by industry and year.

*** indicates significance at the 0.01 level for a two-tailed test.